(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 May 2003 (15.05.2003)

PCT

(10) International Publication Number
WO 03/041342 A1

(51) International Patent Classification⁷: H04L 12/28, H04J 3/14, G01R 31/08, G06F 15/16

(21) International Application Number: PCT/US02/35158

(22) International Filing Date: 1 November 2002 (01.11.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/350,186	2 November 2001 (02.11.2001)	US
10/013,809	7 December 2001 (07.12.2001)	US

(71) Applicant: NETVMG, INC. [US/US]; 47529 Fremont Boulevard, Fremont, CA 94538 (US).

(72) Inventors: KLINKER, Eric; 201 Fourth Street, #511, Oakland, CA 94607 (US). JOHNSON, Jeremy; 3913 Cerrito Avenue, Oakland, CA 94611 (US).

(74) Agents: BACKUS, Kenneth et al.; 2225 E. Bayshore Road, Suite 200, Palo Alto, CA 94303 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

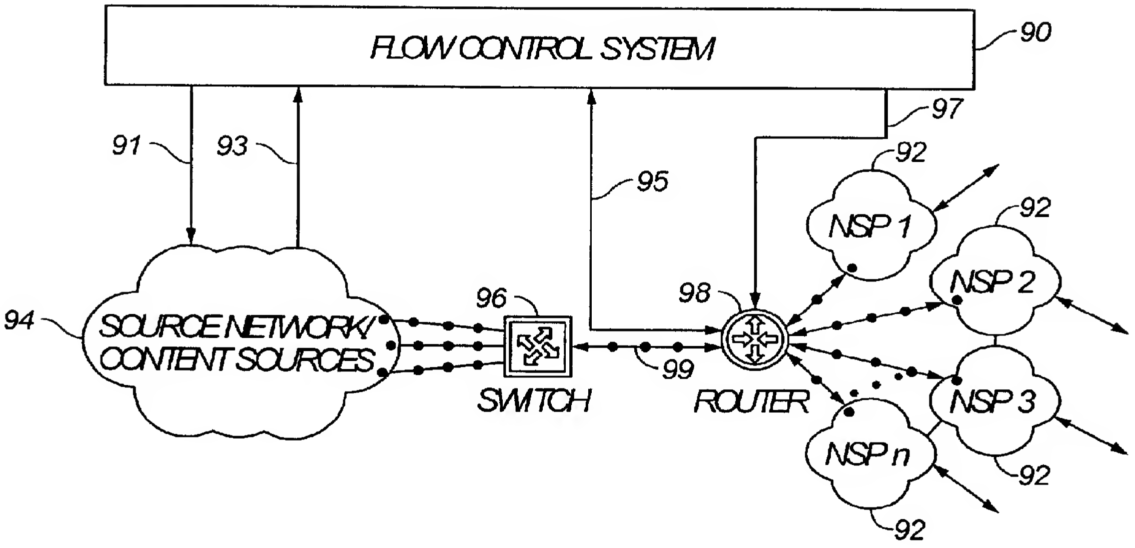
(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEM AND METHOD TO PROVIDE ROUTING CONTROL OF INFORMATION OVER DATA NETWORKS



(57) **Abstract:** A system and a method for controlling routing of data over multiple networks. Accordingly, network users can define specific flow polices (90) to ensure that a particular flow of data traffic (95) maintains an acceptable level of performance, such as in terms of latency, loss, jitter, or an acceptable level usage that includes cost and bandwidth management across multiple networks.

WO 03/041342 A1

System and Method to Provide Routing Control of Information over Data Networks

RELATED APPLICATIONS

This application claims priority from a U.S. Provision Patent Application entitled "SYSTEM AND METHOD TO ASSURE NETWORK SERVICE LEVELS AND BANDWIDTH MANAGEMENT WITH INTELLIGENT ROUTING," identified by Attorney Docket No. 021089-000200US and filed on November 2, 2001, and is incorporated by reference for all purposes. Moreover, U.S. Patent Application entitled "SYSTEM AND METHOD TO ASSURE NETWORK SERVICE LEVELS WITH INTELLIGENT ROUTING," having U.S. Patent Application No. 09/833,219 and Attorney Docket No. 021089-000100US, and filed on April 10, 2001, is incorporated by reference for all purposes.

BACKGROUND OF THE INVENTION

[0001] The present invention relates generally to routing of data over networked communication systems, and more specifically to controlled routing of data over networks, such as Internet Protocol ("IP") networks or the Internet.

[0002] One such data network is the Internet, which is increasingly being used as a method of transport for communication between companies and consumers. Performance bottlenecks have emerged over time, limiting the usefulness of the Internet infrastructure for business-critical applications. These bottlenecks occur typically at distinct places along the many network paths to a destination from a source. Each distinct bottleneck requires a unique solution.

[0003] The "last mile" bottleneck has received the most attention over the past few years and can be defined as bandwidth that connects end-users to the Internet. Solutions such as xDSL and Cable Internet access have emerged to dramatically improve last mile performance. The "first mile" bottleneck is the network segment where content is hosted on Web servers. First mile access has improved, for example, through the use of more powerful Web servers, higher speed communications channels between servers and storage, and load balancing techniques.

[0004] The "middle mile," however, is the last bottleneck to be addressed in the area of Internet routing and the most problematic under conventional approaches to resolving such bottlenecks. The "middle mile," or core of the Internet, is composed of large backbone networks and "peering points" where these networks are joined together. Since peering points have been under-built structurally, they tend to be areas of congestion of data traffic. Generally no incentives exist for backbone network providers to cooperate to alleviate such congestion. Given that over about 95% of all Internet traffic passes through multiple networks operated by network service providers, just increasing core bandwidth and introducing optical peering, for example, will not provide adequate solutions to these problems.

[0005] Peering is when two Network Service Providers ("NSPs"), or alternatively two Internet Service Providers ("ISPs"), connect in a settlement-free manner and exchange routes between their subsystems. For example, if NSP1 peers with NSP2 then NSP1 will advertise only routes reachable within NSP1 to NSP2 and vice versa. This differs from transit connections where full Internet routing tables are exchanged. An additional difference is that transit connections are generally paid connections while peering points are generally settlement-free. That is, each side pays for the circuit or routes costs to the peering point, but not beyond. Although a hybrid of peering and transit circuits (i.e., paid-peering) exist, only a subset of full routing tables are sent and traffic sent into a paid-peering point is received as a "no change." Such a response hinders effective route control.

[0006] Routes received through peering points are one Autonomous System ("AS") away from a Border Gateway Protocol ("BGP") routing perspective. That makes them highly preferred by the protocol (and by the provider as well since those connections are cost free). However, when there are capacity problems at a peering point and performance through it suffers, traffic associated with BGP still prefers the problematic peering point and thus, the end-to-end performance of all data traffic will suffer.

[0007] Structurally, the Internet and its peering points include a series of interconnected network service providers. These network service providers typically maintain a guaranteed performance or service level within their autonomous system (AS). Guaranteed performance is typically specified in a service level agreement ("SLA") between a network service provider and a user. The service level agreement obligates the provider to maintain a minimum level of network performance over its network. The provider, however, makes no such guarantee with other network service providers outside their system. That is, there are no such agreements offered across peering points that link network service providers. Therefore, neither party is obligated to maintain access or a minimum level of service across its peering points with other network service providers. Invariably, data traffic becomes congested at these peering points. Thus, the Internet path from end-to-end is generally unmanaged. This makes the Internet unreliable as a data transport mechanism for mission-critical applications. Moreover, other factors exacerbate congestion such as line cuts, planned outages (e.g., for scheduled maintenance and upgrade operations), equipment failures, power outages, route flapping and numerous other phenomena.

[0008] Conventionally, several network service providers attempt to improve the general unreliability of the Internet by using a "Private-NAP" service between major network service providers. This solution, however, is incapable of maintaining service level commitments outside or downstream of those providers. In addition the common technological approach in use to select an optimal path is susceptible to multi-path (e.g., ECMP) in downstream providers. The conventional technology thus cannot detect or avoid problems in real time, or near real time.

[0009] Additionally, the conventional network technology or routing control technology operates on only egress traffic (i.e., outbound). Ingress traffic (i.e., inbound) of the network, however, is difficult to control. This makes most network technology and routing control systems ineffective for applications that are in general bi-directional in nature. This includes most voice, VPN, ASP and other business applications in use on the Internet today. Such business applications include time-sensitive financial services, streaming of on-line audio and video content, as well as many other types of applications. These shortcomings prevent any kind of assurance across multiple providers that performance will be either maintained or optimized or that costs will be minimized on end-to-end data traffic such as on the Internet.

[0010] In some common approaches, it is possible to determine the service levels being offered by a particular network service provider. This technology includes at least two types. First is near real time active calibration of the data path, using tools such as ICMP, traceroute, Sting, and vendors or service providers such as CQOS, Inc., and Keynote, Inc. Another traditional approach is real time passive analysis of the traffic being sent and received, utilizing such tools as TCPdump, and vendors such as Network Associates, Inc., Narus, Inc., Brix, Inc., and P-cube, Inc.

[0011] These conventional technological approaches, however, only determine whether a service level agreement is being violated or when network performance in general is degraded. None of the approaches to conventional Internet routing offer either effective routing control across data networks or visibility into the network beyond a point of analysis. Although such service level analysis is a necessary part of service level assurance, alone it is insufficient to guarantee SLA performance or cost. Thus, the common approaches fail to either detect or to optimally avoid Internet problems such as chronic web site outages, poor download speeds, jittery video, and fuzzy audio.

[0012] To overcome the drawbacks of the above-mentioned route control techniques, many users of data networks, such as the Internet, use two or more data network connections. Multiple connections increase the bandwidth or throughput of the amount of data capable of traversing the network. With increased bandwidth, performance and reliability of Internet traffic is improved. Also known in the art as "multi-homing," these multiple connections to the Internet generally are across several different network service providers. Multi-homing typically uses Border Gateway Protocol to direct traffic across one or more network service providers' links. Although this traditional approach improves reliability, performance in terms of packet loss, latency and jitter remains unpredictable. The unpredictability arises due to the inherent nature of BGP to not reroute traffic as performance degrades over a particular end-to-end path. Furthermore, BGP tends to direct traffic onto links that only provide the fewest number of hops to the destination, which typically are not the most cost-effective links. This often leads to in efficient routing control techniques, such as over-provisioning of bandwidth across several providers. This, however, leads to increased costs either monetarily or otherwise.

[0013] Given the unpredictability of conventional multi-homing techniques, the network service providers typically deliver unpredictable levels of Internet performance and at different cost structures. No system available today allows Internet customers to manage the bandwidth across multiple providers in terms of at least cost, bandwidth, performance, etc.

BRIEF SUMMARY OF THE INVENTION

[0014] Therefore, there is a need in the art for a system and a method to overcome the above-described shortcomings of the conventional approaches and to effectively and efficiently control routing of data over multiple networks. Accordingly, there is a need to provide intelligent routing control to network users, such as Internet users, to ensure that a particular path used to transport data is selected such that the particular path maintains at least an acceptable level of performance and cost across multiple networks.

[0015] In one embodiment, an exemplary flow control system and method according to one embodiment of the present invention includes one or more modules deployed, for example, at a data network's edge. The flow control system is designed to continuously monitor and route, or re-route, traffic over high-performing paths in real- or near-real-time, thus enabling predictable performance consistent with business-specific application requirements.

[0016] The exemplary system allows the definition and implementation of customer-defined bandwidth usage policies in addition to the definition and implementation of customer-defined performance policies. The user-defined policies enable cost-effective use of existing bandwidth without expensive over-provisioning of network resources. In another embodiment, the system and method provides reports and tools to proactively manage network configurations, such as BGP, and match network performance and cost objectives to the usage of an IP infrastructure.

[0017] In another embodiment, the present invention provides for the monitoring of traffic performance statistics across different network providers, such as Internet transit providers, using multiple techniques. The system is provided information that indicates the destinations where a user's traffic is flowing to and from, the paths being used to reach those destinations, whether the loss or latency performance and transit usage of cost policies that the has defined are being met, and the like. Additionally, the flow control system provides an application-independent traffic flow identification and performance measurement of the traffic flow, accurate measurement of actual end-to-end flow performance across multiple networks from the user's vantage point, real- or near-real time statistics collection. In yet another embodiment, the system continuously detects violations to a user's traffic routing or flow policy for specific destinations, and directs traffic to an alternative path by issuing BGP route updates to a user's router, for example.

[0018] In a specific embodiment, the present invention provides a method of enforcing a policy for data communicated over data networks. Data networks are designed to route data between a first point and a second point, such as between a source and a destination. The first point is coupled to a first network, and in turn, the first network is coupled to one or more second networks. One of the second networks is coupled to the second point for transporting the data communicated to the second point. Each network includes a segment of a path where a path or a path segment includes data flowing, or routing of data, from the first point to the second point. At least two of the networks are coupled at an interconnection point and the data flows through the interconnection point. The method includes monitoring at least one usage characteristic associated with at least one segment, and comparing the at least one usage characteristic with an associated usage requirement of a policy. In another specific embodiment, the method further includes determining if the at least one usage characteristic associated with the routing of data in the first network violates the usage requirement.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1A is an exemplary computer system for presenting to a user a user interface suitable to practice an embodiment of the present invention;

Figure 1B shows basic subsystems in the computer system of Figure 1A;

Figure 1C is a generalized diagram of one exemplary computer network suitable for use with the present invention;

Figure 1D depicts a typical data network using multi- path;

Figure 1E illustrates a simplified data network and flow control system in accordance with a specific embodiment of the present invention;

Figure 2 is a simplified block diagram of one embodiment of a flow control system according to one embodiment the present invention;

Figure 3 is a functional block diagram of an exemplary passive calibrator of Figure 2;

Figure 4 is a functional block diagram of an exemplary content flow analyzer of Figure 3;

Figure 5 is a functional block diagram of an export flow analyzer of Figure 3 in accordance to one embodiment of the present invention;

Figure 6 is a functional block diagram of a passive flow analyzer of Figure 3 according to a specific embodiment;

Figure 7 is a simplified timing diagram of determining network performance metrics with an exemplary flow control system located near a client or a source;

Figure 8 is a simplified timing diagram of determining network performance metrics with an exemplary flow control system located near a server or a destination;

Figure 9 is a network diagram of an exemplary passive calibrator with distributed packet capture according to another embodiment of the present invention;

Figure 10 is a network diagram of distributed passive flow elements according to yet another embodiment of the present invention;

Figure 11 is a functional block diagram of the distributed passive flow elements of Figure 10 according to still yet another embodiment of the present invention;

Figure 12 is a detailed block diagram of an exemplary usage collector according to a specific embodiment of the present invention;

Figure 13 is a block diagram of a route server using an associated configuration element receiving either multiple BGP4 feeds or at least one iBGP feed according to one embodiment of the present invention;

Figure 14 is a graphical representation illustrating an exemplary method to determine the amount of bandwidth available that can be used without additional cost in accordance to the present invention;

Figure 15 is a graphical representation illustrating an exemplary method to calculate billable rates in accordance to the present invention;

Figure 16 is a graphical representation depicting an exemplary method to calculate billable rates using short range forecasting in accordance to the present invention; and

Figure 17 is a representation of an exemplary address or prefix list according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE SPECIFIC EMBODIMENTS

[0019] Detailed descriptions of the embodiments are provided herein. It is to be understood, however, that the present invention may be embodied in various forms. Therefore, specific details disclosed herein are not to be interpreted as limiting, but rather as a basis for the claims and as a representative basis for teaching one skilled in the art to employ the present invention in virtually any appropriately detailed system, structure, method, process or manner.

[0020] Figures 1A, 1B, and 1C illustrate basic hardware components suitable for practicing a specific embodiment of the present invention. Figure 1A is an illustration of an exemplary computer system 1 including display 3 having display screen 5. Cabinet 7 houses standard computer components such as a disk drive, CD-ROM drive, display adapter, network card, random access memory (RAM), central processing unit (CPU), and other components, subsystems and devices. User input devices such as mouse 11 having buttons 13, and keyboard 9 are shown. Other user input devices such as a trackball, touch-screen, digitizing tablet, voice or visual recognition, etc. can be used. In general, the computer system is illustrative of but one type of computer system, such as a desktop computer, suitable for use with the present invention. Computers can be configured with many different hardware components and can be made in many dimensions and styles (e.g., laptop, palmtop, pentop, server, workstation, mainframe). Any hardware platform suitable for performing the processing described herein is suitable for use with the present invention.

[0021] Figure 1B illustrates subsystems that might typically be found in a computer such as computer 1. In Figure 1B, subsystems within box 20 are directly interfaced to internal bus 22. Such subsystems typically are contained within the computer system such as within cabinet 7 of Figure 1A. Subsystems include input/output (I/O) controller 24, System Memory (or random access memory "RAM") 26, central processing unit CPU 28, Display Adapter 30, Serial Port 40, Fixed Disk 42, Network Interface Adapter 44 (e.g., Network Interface Card, or NIC), which in turn is configured to communicate with a network, such as by electrical, radio, or optical means known in the art. The use of bus 22 allows each of the subsystems to transfer data among subsystems and, most importantly, with the CPU, where the CPU might be a Sparc™, an Intel CPU, a PowerPC™, or the equivalent. External devices can communicate with the CPU or other subsystems via bus 22 by interfacing with a subsystem on the bus. Thus, Monitor 46 connects with Display Adapter 30, a relative pointing device (e.g. a mouse) connects through a port, such as Serial Port 40. Some devices such as Keyboard 50 can communicate with the CPU by direct means without using the main data bus as, for example, via an interrupt controller and associated registers.

[0022] As with the external physical configuration shown in Figure 1A, many subsystem configurations are possible. Figure 1B is illustrative of but one suitable configuration. Subsystems, components or devices other than those shown in Figure 1B can be added. A suitable computer system also can be achieved using fewer than all of the sub-systems shown in Figure 1B. For example, a standalone computer need not be coupled to a network so Network Interface 44 would not be required. Other subsystems such as a CD-ROM drive, graphics accelerator, etc. can be included in the configuration without affecting the performance of the system of the present invention.

[0023] Figure 1C is a generalized diagram of a typical network that might be used to practice an embodiment of the present invention. In Figure 1C, network system 80 includes several local networks coupled to computer data network 82, such as the Internet, WAN (Wide Area Network), or similar networks. Network systems as described herein refer to one or more local networks and network service providers that make up one or more paths from a source to a destination and visa versa. Network systems, however, should be understood to also denote data networks that including one or more computing devices in communication using any networking technology. Although specific network protocols, physical layers, topologies, and other network properties are presented herein, the present invention is suitable for use with any path-diverse network (e.g., a multi-homed network interconnected to other networks), especially those networks that employ Internet Protocol (IP) for routing data, such as flows having one or more packets of information according to the protocol. Furthermore, although a specific implementation is not shown in Figure 1C, one having ordinary skill in the art should appreciate that a flow control system according to the present invention can be deployed within one or more data networks 82 or configured to operate with network system 80.

[0024] In Figure 1C, computer USER1 is connected to Server1, wherein the connection can be by any network protocol, such as Ethernet, Asynchronous Transfer Mode, IEEE standard 1553 bus, modem connection, Universal Serial Bus, etc. The communication link need not be a wire but can be infrared, radio wave transmission, etc. As depicted, Server1 is coupled to the data network 82, such as the Internet or, for example, any other data network that uses Internet Protocol for data communication. The data network is shown symbolically as a collection of server routers 82.

[0025] The exemplary use of the Internet for distribution or communication of information is not strictly necessary to practice the present invention but rather is merely used to illustrate a specific embodiment. Further, the use of server computers and the designation of server and client machines are not crucial to an implementation of the present invention. USER1 Computer can be connected directly to the Internet. Server1's connection to the Internet is typically by a relatively high bandwidth transmission medium such as a T1 line, a T3 line, Metro Area Ethernet, or the like, although it might be connected in a similar fashion as with USER1. Similarly, other computers 84 are shown utilizing a local network (e.g., Local Area Network, or LAN) at a different location from USER1 Computer. The computers at 84 are coupled via Server2 to the Internet. Although computers 84 are shown to include only a single server (e.g., Server2), two or more servers can be connected to the local network associated with computers 84. The USER3 and Server3 configuration represent yet a third network of computing devices.

[0026] Figure 1D shows the effects of typical multi-path (e.g., ECMP) techniques on a route control system using active calibration alone. Two possible paths exist between Washington DC and San Jose for a given network service provider. The first path 170 traverses New York, Chicago and Seattle. The second path 171 traverses Atlanta, Dallas and Los Angeles. Suppose that the cost of using either of the paths is equal in the routing protocol. Most router vendors, when presented with two equal costs paths, will load share traffic between them making sure that paths in the same flow will follow the same route. The path selection process is vendor-specific and generally relies on known source and destination IP addresses. Unless the source IP address and destination IP address are the same, the traffic may take a different equal-cost path. The implications for path calibration are that the active probes sent across the network between Washington DC and San Jose may take the northern path through Chicago 172 while the customer's traffic may take the southern path through Dallas 173, because while the destination IP address is the same, the source IP address is different. Thus, the path measured may not be the path that is actually taken by the customer's traffic. The present invention, among other things, intelligently controlled routes containing data traffic using a system and a technique to assure service levels of customer data traffic in accordance with the present invention.

[0027] Figure 1E illustrates an exemplary data network within a portion of a network system 80 of Figure 1C including NSPs 92, and a flow control system in accordance with a specific embodiment of the present invention. Exemplary flow control system 90 is configured to communicate with one or more network elements of the data network. Although flow control system 90 is shown external of and in communication with the elements of source network 94, switch 96, and router 99, flow control system 90 can be wholly embodied in any of the elements shown, or alternatively, can be distributed, in portions, over each of the elements. In another embodiment, flow control system 90 resides on one or more servers or network elements within exemplary source network 94.

[0028] An exemplary data network includes one or more source networks 94. A source network 94 typically is a local network including one or more servers owned and operated by application service providers, managed service providers, content delivery networks, web hosting companies, individual enterprises, corporations, entities and the like.

Such service providers typically communicate information to users that are further removed from the multi-homed network service providers 92, such as NSP 1, NSP 2, NSP 3, . . . and NSPn. In one example, network service providers 92 are coupled to a source network or source point as to be considered a first set of data networks. These NSPs, or first set of data networks, are in turn coupled to a second set of networks, wherein the second set is connected to multiple other networks, thus establishing one or more paths from a source to a destination. A path as describe herein can be a route from a source to a destination that is divided into segments, each segment residing wholly within a provider.

[0029] The multiple connections between router 98 and multiple network service providers 92 provide an operator of source network 94 to direct data traffic according to the best performing network service provider. Switch 96 operates to transfer bi-directional data 99, such as IP data, bi-directionally from source network 94 to router 98. Although a single router and switch are shown, one having ordinary skill in the art will appreciate that either additional routers and switches or other suitable devices can be substituted according to another embodiment of the present invention. Moreover, switch 96 need not be used to practice the subject invention. In a specific embodiment, router 98 includes one or more routers running an exemplary protocol, such as Border Gateway Protocol (e.g., BGP4, such as Cisco™ or Juniper implementations™), for example, and preferably has route visibility across multiple network service providers.

[0030] In an embodiment of flow control system 90, system 90 operates to measure end-to-end (i.e., source to destination and destination to source) data traffic 95 in terms of flow characteristics, such as performance, cost, bandwidth, and the like. Flow control system 90 also generates statistics associated with data paths across multiple network service providers in real time, or near-real time. Such statistics are communicated to source network 94 for providing network engineering personnel, for example, with report information 91 such that on-the-fly reports are created to provide information related to route-change activity, traffic performance as delivered to selected destinations and transit provider usage (i.e., bandwidth), cost, and the like.

[0031] In one embodiment of the present invention, a local computing device uses report information 91 from system 90 to generate visual and graphical representations on, for example, a user-friendly interface ("UI") where the representations are indicative of data traffic along one or more paths (e.g., paths between a source and a destination). Network personnel, or any entity responsible with flow control, with access to source network 94 then can provide control information 93 to flow control system 90 to modify system operation by, for example, changing data traffic flow from a under-performing current, or default, path to a better performing path. Intervention by network personnel, however, is not necessary for flow control system 90 to operate in accordance with the present invention.

[0032] Flow control system 90 further functions to compare specific data traffic flows (i.e., both uni- and bi-directional traffic flows outbound from and inbound into the data network) to determine whether a particular traffic flow meets one or more rules of an associated flow policy. A flow policy, as referred to herein, includes a set of one or more rules that is associated with a particular data traffic flow related to particular system user (e.g., as denoted by IP address prefix).

[0033] A rule, or criterion, is a minimum level, a maximum level or a range of values that defines acceptable routing behavior of an associated with a traffic flow characteristic. For example, a rule can set: the maximum acceptable cost, with or without regard to network service provider cost; the maximum load or bandwidth usage associated with traffic flows through specific providers; a range of acceptable (or non-acceptable) service providers; the maximum acceptable latency or loss over one or more paths across multiple network service providers; acceptable ranges of performance for each network service provider,.such as maximum burst limits, minimum performance commitments and range of costs (i.e., cost structures with regards to time of day, type of traffic, etc.); and any other data flow characteristic that can influence the measurement or the control of data traffic.

[0034] Flow control system 90 further operates to detect when one or more rules, or flow policies, are violated and then to take remedial action. That is, flow control system 90 enforces policies associated with data traffic flow by correcting detrimental deviations in performance (i.e., service level assurance), costs or bandwidth (i.e., load in terms of percent capacity available per path). Flow control system 90 makes such corrections based on real- or near-real time traffic analysis, local path diversity (i.e., modifying one or more egress paths from a data network), and visibility into downstream available paths. For example, for a destination related to a specific traffic flow, flow control system 90 directs, or re-directs, traffic to one or more alternative paths to resolve a particular flow's deviation in terms of flow characteristics, from its flow policy.

[0035] Figure 2 illustrates a specific embodiment of flow control system 90 of Figure 1D. In another embodiment, flow control system in figure 2 is a reactive flow control system. That is, a reactive flow control system is designed to react to policy violations indicating sub-standard routing of data traffic over one or more data networks or service providers (i.e., addresses pass-fail criteria) rather than optimizing performance at some targeted level of acceptable operation.

[0036] Flow control system 200 includes controller 205, passive calibrator 203, active calibrator 208, configuration element 211, and usage collector 214, each of which can be realized in hardware, software, or a combination thereof. For example, controller 205, passive calibrator 203, active calibrator 208, configuration element 211, and usage collector 214 are software modules designed to perform specific processes, as described herein, in accordance to the present invention. Such modules can reside in one or more computing devices, such as the computing devices shown in Figure 1A, or alternatively, over one or more USER-type machines (i.e., servers) coupled over a data network or network system.

[0037] Exemplary passive calibrator 203, active calibrator 208 and usage collector 214 are coupled to controller 205 to, in part, provide flow characteristics of data traffic. Controller 205 receives monitored flow characteristics as well as flow policies to be enforced. Controller 205 is configured to determine if a flow policy is violated, and upon detection of such a violation, then to select a remedial action to resolve the violation. Configuration element 211 is coupled to controller 205 used to receive information to initiate remedial actions and is configured to communicate such actions to data director 220. Thereafter, data director 220 implements the corrective action to resolve the pending violation, for example, by changing the traffic flow from the current path to a better performing path.

[0038] Additionally, flow control system 200 includes traffic repository 221 and flow policy repository 218. Exemplary traffic repository 221 and flow policy repository 218 are databases, such as a storage device, configured to store a large number of records in one or more data structures. Traffic repository 221 is designed to store and to communicate information related to traffic and route characteristics, and flow policy repository 218 is designed to store and to communicate policy information or rules to govern the performance and cost of each of the data traffic flows. One having ordinary skill in the art of database management should appreciate that many database techniques may be employed to effectuate the repositories of the present invention.

[0039] In operation, flow control system 200 of Figure 2 monitors egress and ingress data flow 201, such as IP data traffic, to determine whether data flow 201 to and from source network is within the performance tolerances set by the associated flow policy. Flow control system 200, in one embodiment, receives data flow 201 by replication, such as by a network switch, by using a splitter, such as an optical splitter, or any other tapping means know to those having ordinary skill in the art. Data flow 202, which is exactly, or near exactly, the same as the information contained within data flow 201, is provided to passive calibrator 203.

[0040] Passive calibrator 203 monitors the data traffic of data flow 201 and communicates information 204 related to the traffic and traffic performance to controller 205. Controller 205 is configured to receive policy data 206 representing one or more policies that correspond to a particular traffic flow, such as a particular data flow. Moreover, the particular data flow can be associated with a certain user identified by a destination prefix, for example. From policy data 206, controller 205 determines the levels of performance, cost, or utilization that the particular traffic is to meet. For example, controller 205 determines whether a particular traffic flow of data flow 201 is meeting defined performance levels (i.e., service levels) as defined by one or more requirements or criteria, such as inbound and outbound network latency, packet loss, and network jitter.

[0041] Active calibrator 208 functions to send and to receive one or more active probes 207, of varying types, into and from the data networks. These probes are designed to measure network performance including, path taken across one or more available providers (i.e., to determine if a provider is a transit AS rather than peer AS), next hop-in-use, and other network parameters. To activate active calibrator 208, controller 205 sends an active probe request 209 to active calibrator 208. Such a request is required if controller 205 determines that additional information regarding alternative paths or network system characteristics are necessary to better enforce policies in reactive flow control systems, or alternatively, to prevent such policy violations optimized flow control systems.

[0042] Usage collector 214 is configured to receive NSP data 217 representing one or more network provider configurations. Generally, such configurations include the number of paths ("pipes") associated with each provider and the size thereof. Additionally, NSP data 217 can relate to a provider's cost or billing structure and can also include each provider's associated set or sub-set of addresses, each provider's billing methods (i.e., byte/min, etc.), etc. Moreover, usage collector 214 is configured to collect usage information 213 from the network elements, such as switches, border routers, provider gear, and other devices used to transport data over data networks. Usage collector 214 is configured to provide controller 205 with provider utilization and billing information 215, which represents aggregated data based upon NSP data 217 and usage information 213. Utilization and billing information 215 includes data that represents cost, billing, utilization, etc., for each network service provider of interest.

[0043] One having ordinary skill in the art should appreciate that NSP data 217 can be provided to usage collector 214 in a variety of ways. For example, the data can be provided the data paths used by the data flows or can be provided by an entity having authority to do so, such a network engineer entering the data into a computing device in source network 94 of Figure 1E.

[0044] Moreover, usage collector 214 is configured to monitor usage characteristics defining a network service provider's data traffic capacity, costs, etc. Usage information 213 provided to usage collector 214 includes usage characteristics from network elements, such as switches, border routers, routers, provider gear, and other devices used to transport data over data networks. Usage refers to the data (i.e., raw data such as X Mb samples at time(0)) that represents instantaneous or near instantaneous measurement of characteristics (i.e., usage characteristics) that define, for example, the load and available capacity of each network service provider. Utilization is the usage rate over time. For example, suppose the usage collector monitoring NSP1 measures its utilization, or capacity over time, as X Mb at time(0) and Y Mb at time(1). This raw data, or usage, is used to calculate utilization, or usage rate for NSP1 (e.g., Y-X/ time(1)-time(0)). Bandwidth is the total capacity each path or segment of path available for traffic flow. In one embodiment, the usage can be measured in any segment in any path at any number of hops or networks from a first point. Load is typically defines the amount of capacity a particular path is used to carry data traffic and can be expressed as load/bandwidth.

[0045] Usage collector 214 is designed to generate utilization and billing information 215 based upon usage information 1213 and NSP data 217. Since each of the providers has different cost and billing structures, as well as methods of determining usage costs, usage collector 214 operates to aggregate usage information 213 accordingly to provide controller 205 with utilization and billing information 215.

[0046] Usage collector 214 then provides the utilization billing information 215 to controller 205 for each network service provider of interest. One having ordinary skill in the art should appreciate that the usage collector can provide additional information based upon the provider usage information, to the controller, as needed to better effectuate route control.

[0047] Controller 205 collects information (i.e., aggregated performance and usage characteristics) from each of passive calibrator 203, active calibrator 208, usage collector 214, and optionally traffic repository 221. Based upon the information collected, controller 205 determines a course of action that best alleviates the policy violations in respect to the information represented by policy data 206 that is conveyed to controller 205. Once the coarse of action is determined, controller 205 initiates and sends a network routing change request 212 to configuration element 211. In a specific embodiment, controller 205 also provides data representing one or more alternate data paths that can be used resolve the policy violation.

[0048] Configuration element 211 is designed to communicate routing changes in the network to data director 220. Once configuration element 211 sends one or more routing changes, data director 220 then moves data flow 201 from a current path to another path (e.g., from NSP1 to NSPn or a first path of NSPI to a second path of NSPI). Data director 220 thus operates to distribute traffic to these destinations across multiple network service provider links based on, for example, the cost and performance measured across each link.

[0049] In operation, configuration element 211 communicates one or more routing changes 210 with data director 220, for example, by using a routing protocol such as BGP. Configuration element 211 functions to dynamically control routing behavior by modifying the source address of the traffic passing through configuration element 211. The source address is modified in a way that improves application performance as well as cost requirements.

[0050] The following discussion is a more description of each of the elements of an exemplary control system 200. Referring back to active calibrator 208, active calibrator 208 provides active mechanisms within system 200 for determining the nature of downstream or upstream paths. This information is typically not available in any conventional protocol used on data networks such as the Internet, and must be collected external to the normal processes networking. As shown in Figure 2, active calibrator 208 is coupled to controller 205 to provide at least a destination prefix that is not meeting the policy requirements, such as minimum performance level. Once received, active calibrator 208 then initiates a calibration process that determines most or all of the available network paths to the destination address as well as performance levels. Controller 205 is designed to select the most suitable probes that active calibrator 208 is to use, based on the particular policy requiring enforcement or correction, and thereafter to initiate active probing of network paths using active calibrator 208.

[0051] In one embodiment, active calibration probes are communicated to available network or Internet paths via probe path 207 . The returning active calibration probes enter via probe path 207 into active calibrator 208. Active calibrator then forwards probe information 209 to controller 205, which contains performance information including alternate available paths. Controller 205 then determines how best to enforce the specifics of the policy associated with the subject traffic flow. Exemplary active calibrator 208 employs active calibration mechanisms to provide, for example, long term statistics.

[0052] In another embodiment of the present invention, active calibrator 208 resides in data director 220 within, or alternatively, can be integrated into controller 205. There are several proprietary implementations of commercially available routers suitable to practice the present invention. One example of suitable active probes is the RMON probe. Cisco systems use Service Assurance Agent ("SAA") that is derived from the remote monitoring ("RMON") probes to send out active probes. SAA allows routers to measure and report network-originated application round trip times. Although not every probe mentioned below is available in SAA for network calibration, one skilled in the art would appreciate how each of the following might be implemented to practice one or more embodiments of the present invention.

[0053] An exemplary active calibrator 208 can use ICMP (Internet Control Message Protocol) echo request or other ping-type probes, lightweight TCP-based probes, Sting probes, "pathchar" probes, lightweight probes using User Datagram Protocol ("UDP") packets with a predefined TTL (time to live), traceroute probes, or other active that are suitable for use by active calibrator 208 in accordance with the present invention.

[0054] These probes are received back by active calibrator 208 of Figure 2 are sent out by their source addresses. Such probes are all sourced and received on an exemplary stats computer system resident, for example, in the local premises, or as a stats process on a router. In another embodiment, active calibrator and the of its use of probes operate in accordance to probes described in a U.S. Patent Application, entitled "System and Method to Assure Network Service Levels with Intelligent Routing," having U.S. Pat. Application No.

09/833,219 and Attorney Docket No. 021089-000100US and filed on April 10, 2001, and is incorporated by reference for all purposes.

[0055] Exemplary passive calibrator 203 of Figure 2 is configured to receive, without interfering with, network communication data 201, such as customer network or Internet traffic. Network communication data path 201 (i.e., IP data traffic), as monitored by passive calibrator 203, includes the default or currently routed path of the data traffic that is and is provided to passive calibration element 203 from data director 220. The currently routed path is, for example, the path (e.g., hop-by-hop) between routers that a packet would take, as determined by standard routing protocols. Passive calibrator 203 is coupled (i.e., electrically, optically, by radio waves, etc.) to controller 205 to provide information which indicates whether the specific IP data traffic is within the range of acceptable performance metrics, such as determined by a flow policy. Passive calibrator 203 operates to instantaneously monitor all traffic received via data flow 202 and is designed to overcome the complications of relying solely on active traffic analysis, such as EMCP, as shown with respect to Figure 1D. When the controller addresses policy violations, for example, passive calibrator 203 operates to overcome the complications of performing only active traffic analysis in the presence of multi-path (e.g., ECMP).

[0056] In another embodiment of the present invention, passive calibrator 203 examines the traffic stream in both directions (i.e., ingress and egress) and classifies each of the traffic streams into flows. Traffic flows, are monitored within passive calibrator 203 according to the underlying protocol state (e.g., such as regarding TCP sessions) over time. For example, passive calibrator 203 classifies the traffic flow according to round trip latency, percentage of packets lost, and jitter for each of the traffic routes or flows. Such traffic route information is used to characterize the "end-to-end" performance of the paths carrying the traffic flows, which includes flow rates, and is aggregated into a series of network prefixes.

[0057] As described above, passive calibrator 203 is coupled to store, fetch and update traffic and route information stored in traffic repository 221 (connection not shown). Exemplary traffic repository 221 is a database configured to store and to maintain data representing traffic and route information that is useful to the end user employing a flow control system, such as system 200 of Figure 2, as well as the operators of, for example, an network service provider. The data within traffic repository 221 includes long term statistics about the traffic. These statistics will be used for reporting, analysis purposes, and providing general feedback to a user of a flow control system according to the present invention.

[0058] Such feedback will consist, for example, of types of traffic being sent, source addresses, destination addresses, applications, traffic sent by ToS or DSCP ("DiffServ Code Point") setting (which might be integrated into a differentiated billing system), and volume of traffic. These statistics are fed into traffic repository 221 where, for example, a reporting engine or some other analysis process has access to them. The information stored in traffic repository 221 is data representing such traffic route characteristics arranged in any suitable data structure as would be appreciated by one skilled in the art.

[0059] Figure 3 is a detailed functional block diagram showing exemplary elements of a passive calibrator 303 according to an embodiment of the present invention. Passive calibrator 303 includes, for example, passive flow analyzer 330, export flow analyzer 331, and content analyzer 332.

[0060] In one embodiment, passive flow analyzer 330 performs passive analysis on the traffic to monitor current traffic flow characteristics so the controller can determine whether the monitored current traffic flow meets associated policy requirements. Export flow analyzer 331 performs passive analysis on exported flow records from a network device, such as from those devices (e.g., router) that advertise traffic type, source and destination addresses, and other information related to the traffic that it travels across service provider links. An example of such a network device is Cisco's Netflow™ product. In another embodiment, passive flow analyzer 330 operates in accordance to the passive flow analyzer described in the above-mentioned U.S. Patent Application No. 09/833,219.

[0061] Content Flow Analyzer 332 performs passive analysis of specific elements of data content, such as web site content. Export flow analyzer 331 and content flow analyzer 332 determine a set of relevant prefixes or a prefix list 334 that is associated with a specific user's policy. Prefix list 334 is sent as data representing such prefixes to an active detection process in the controller. Prefix list 334 can be one or more lists or data structures configured to store data representing performance and usage characteristics and are designed to be receive a query, for example, by the controller. Once queried, the passive flow analyzer provides the one or more prefix lists, or portions thereof, to the controller for use in determining a policy violation, for determining which routes or path comply with the flow policy, which path is the optimum path for routing data, and the like. An exemplary prefix list that can be generated by export flow analyzer 331 and content flow analyzer 332, as well as passive flow analyzer 330.

[0062] Figure 17 shows an exemplary data structure 1900 suitable for providing for one or more of the prefix lists described herein. Data structure, or list, 1900 includes many IP addresses 1920 with many records 1910 associated with each address (e.g., destination) or prefix of variable granularity. Each record 1910 includes an address 1920 (or prefix), a number of occurrences during a time period 1930, number of bytes sampled 1940, time interval in which sampling occurred (delta t) 1950, new prefix flag 1960 (1 represents new prefix, 0 represents old prefix), or the like.

[0063] List 1970 includes aggregate flow information for each address 1920 or prefix. For example, record 1975 includes the following data: for address 1.2.4.7, this address was monitored four times during the sampling time interval (delta)t with a total flow volume of

360 bytes. With record 1990 having a new prefix flag set (i.e., first time this address has been monitored), new prefix list 1980 includes address 1.2.4.9 having one occurrence (first time) over (delta)t interval. One having ordinary skill in the art should appreciate that other relevant data may be monitored and can be stored in list 1900. Moreover, the data representing address, occurrence, number of bytes, time interval, etc., can be used to manipulate the data such in a way that the controller can easily obtain.

[0064] For example, the data stored within a list 1920 can be aggregated or grouped according to address or prefix. As shown in Figure 17, aggregate list 1995 includes the group of addresses corresponding to 1.2.4.X. For example, the record 1997 of aggregate addresses contains data indicating that the aggregation of addresses had been monitored five times during the time interval and had a total volume of 540 bytes. One having ordinary skill in the art should appreciate that addresses or prefixes can be grouped or aggregated in many ways.

[0065] Export flow analyzer 331 and content flow analyzer 332 also are configured to notify controller 305 when a previously unseen prefix has been added to the prefix list 334. New prefix notification signal 335 enables the control element 1005 to establish a new baseline performance for this prefix and to seed the routing table with a non-default route, or alternative route (i.e., non-BGP), if necessary. In one embodiment, export flow analyzer 331 and content flow analyzer 332 provide for monitoring of performance characteristics.

[0066] Content flow analyzer 332 is typically used when the main source of traffic flow 340 is web site or other content. Content source 341 can be configured such that special or premium content 342 that must be optimized can be identified by the flow control system by using, for example, an embedded URL 343. URL 343 redirects the client to a small content server running on the content flow analyzer 332. Content flow analyzer 332 receives a request for the small content element, which is generally a small image file (e.g., 1 x 1 GIF) and is invisible or imperceptible in relation with the main original content, and responds to the client with the small content element 344. Content flow analyzer 332 then stores or logs this transaction, and by using these logs, content flow analyzer 332 is able to perform aggregation and assemble content prefix list 334. The list 334 is passed along to controller 205, for example, for active service level monitoring and policy enforcement.

[0067] Figure 4 illustrates a functional block diagram of an exemplary content flow analyzer 432. Content flow analyzer 432 handles requests 420 for a small element of content, which is, for example, a 1x1 pixel image file that is imperceptible (although it need not be) on the resulting page. The small element is associated with the premium or generally specific pages of a larger set of content. The small element is, for example, a small redirect URL embedded within the content.

[0068] The small redirect URL acts to generate an HTTP request 420 in response to the small element of content. Content flow analyzer 432 sees this request 420 and responds 422 to it with, for example, a lightweight HTTP server 453. This server is fast and lightweight, and does nothing other than respond with the image file. The lightweight web server 453 logs the IP address of the client requesting the web page, and sends the one or more addresses to aggregator 454. Aggregator 454 aggregates, or collates, individual IP elements 424 into prefixes of varying granularity (e.g., /8 through /32) and also aggregates the frequency that each prefix is seen over an interval of time.

[0069] That is, aggregator 454 classifies prefixes according to its frequency of occurrence and provides aggregated (i.e., grouped) prefixes 426 to prefix list generator 455. Prefix list generator 455 creates destination prefix list 428 according, for example, to a prefix's importance in relation to the overall operation of the system as defined by the aggregated or grouped prefixes 426. For example, each monitored traffic flow is examined to determine the performance characteristics associated with a destination prefix or address.

[0070] Aggregate prefixes 426 are generally classified in terms of flow frequency, and average or total flow volume. Prefix list generator 455 sends updates to current prefix list 428 to controller 205 of Figure 2, and also notifies other elements of the system with new prefix notification signal 432 when a new prefix is observed. Prefix list generator 455 stores the prefix information 430 to persistent storage for reporting and analysis purposes. A new prefix provides an additional alternate path or path segment that was unknown up until a certain point of time. The new alternate path or path segment associated with the new prefix can provide for flow policy compliance, and thus can have be used to re-route or alter routing of data to obviate a policy violation.

[0071] Referring back to Figure 3, export flow analyzer 331 operates in conjunction with network elements that can export (i.e., communicate) flow information in a format useable by analyzer 331. One exemplary format is the Cisco NetFlow™ export format. Any network element designed to export flow information, such as router 345 or a layer 2 switch, thus is also configured to passively monitor the traffic it is processing and forwards export records 346 to export flow analyzer 331. Export flow analyzer 331 functions to process export flow records 346, aggregates the flows into prefix elements, and generates prefix list 334. The prefix list is generally a subset of all prefixes observed by the flow control system. A prefix is selected from all prefixes based upon flow volume and flow frequency over an observation period. The selected prefix then is placed into prefix list 334 before the list passed along to controller 205 of Figure 2, for example.

[0072] Figure 5 illustrates a functional block diagram of exemplary export flow analyzer 531. Export flow analyzer 531 includes format interpreter 549, parser 550 and prefix list generator 552. Format interpreter 549 is configured to receive export flow datagrams 520 from the network elements designed to send them. Format interpreter 549 then communicates individual flow information 522 to parser 550. Parser 550 operates to interpret destination IP elements from the flows monitored by the passive calibrator. Parser 550 also aggregates traffic flow according to total flow volume or transportation rate (e.g., in bytes/time unit) as well as flow frequency of destination addresses, for example, into aggregate elements. Thereafter, parser 550 sends the aggregate elements 524 to aggregator 551. Aggregator 551 then generates prefix-level destination information 526 (i.e., aggregate prefix volume and frequency) at a variety of prefix granularities (e.g., from /8 up through /32). In other words, aggregator 551 determines the frequency, session, or for a specific prefix and the aggregate volume of occurrences related to that prefix over an observed time interval.

[0073] Destination prefix list 528 is generated by prefix list generator 552 by, for example, ranking and organizing traffic flow characteristics related to prefixes in order of relative importance. List 528 contains data representing an aggregation of prefixes prefix list 528 and is organized in determines the relevance, as determined by the system or an entity to ensure policy enforcement. For example, one or more prefixes can be ordered in terms of flow frequency and average or total flow volume in relation together prefixes available in the overall system. Prefix list generator 552 sends updates to the current prefix list to controller 205 of Figure 2 and also notifies other elements of the system when a new prefix is observed via a new prefix notification signal 532. Prefix list generator 552 stores all prefix information 530 to persistent storage for reporting and analysis purposes.

[0074] Figure 6 illustrates a function block diagram of an exemplary passive flow analyzer 630 of Figure 3. In one embodiment, passive flow analyzer 630 is designed to generate prefix list 634 and new prefix notification signal 635 and generates aggregated flow data 680, including network performance and usage statistics grouped into relevant characteristics. For example, prefixes of a certain size can be aggregated, or grouped, from highest traffic volume to lowest as observed over time. The aggregated flow data 680 is communicated to controller 605 and are used by the controller to determine whether the current traffic flow violates or fails to conform to an associated flow policy for a given destination. The passive flow analyzer 630 also functions to store aggregated flow data 680 in traffic repository 621, where it can be used for characterizing historical route and traffic flow performance. In another embodiment of the present invention, a prefix list generator is not included in the passive flow analyzer of Figure 6.

[0075] Passive Flow Analyzer 630 uses a copy of the traffic 602 via a passive network tap or spanned switch port, as shown in Figure 2, to monitor the network performance for traffic. Passive flow analyzer 630 also can monitor and characterize UDP traffic patterns for detection of anomalous behavior, such as non-periodic traffic flow, or the like. Passive flow analyzer 630 can use various neural network techniques to learn and understand normal UDP behavior for the application in question, and indicate when that behavior has changed, possibly indicating a service level violation which can be verified or explained with well known active probing techniques.

[0076] Additionally, passive flow analyzer 630 is designed to be "application-aware" according how each of the particular traffic flows is classified. Traffic can be classified according to the classifier described in the above-mentioned U.S. Patent Application No. 09/833,219. That it, Passive flow analyzer 630 can inspect the payload of each packet of traffic 602 to interpret the performance and operation of specific network applications, such as capture and interpretation of the Realtime Transport Control Protocol ("RTCP") for voice over IP ("VoiP"), for example.

[0077] In Figure 6, passive flow analyzer 330 includes packet capture engine 650, packet parser 651, correlation engine 652, and aggregator 653. Packet capture engine 650 is a passive receiver configured to receive traffic (e.g., IP data traffic) coming into and out of the network. Capture of traffic is used to facilitate traffic analysis and for determining a whether a current traffic route meets minimum service levels or policy requirements. Packet capture engine 650 is designed to remove one, several or all packets from a traffic stream, including packets leaving the network and entering the network. Packet capture engine 250 operates to remove certain packets up, for example, from the network drivers in the kernel into user space by writing custom network drivers to capture part of a packet. Using DMA, the partial packet can be copied directly into user space without using the computer CPU. Such packets are typically removed according to one or more filters before they are captured. Such filters and the use thereof are well known in the art and can be designed to, for example, remove all types of TCP traffic, a specific address range or ranges, or any combination of source or destination address, protocol, packet size, or data match, etc. Several common libraries exist to perform this function, the most common being "libpcap." Libpcap is a system-independent interface for packet capture written at the Lawrence Berkeley National Laboratory. Berkeley Packet Filter is another example of such capture program.

[0078] Parser 651 is coupled to receive captured raw packets and operates to deconstruct the packets and retrieve specific information about the packet from each in the traffic flow. Exemplary parser 651 extracts information from the IP and TCP headers. Such extracted information from the IP headers include source and destination IP addresses, DSCP information encoded in the ToS (i.e., "type of service") bits, and the like. DSCP carries information about IP packet QoS requirements. Each DSCP defines the Per Hop Behavior of a traffic class. DiffServ has 64 code points so that it can define 64 different types of traffic classifications. TCP header information includes source and destination port numbers, sequence number, ACK number, the TCP flags (SYN, ACK, FIN etc.), the window size, and the like.

[0079] TCP elements parsed from the TCP headers are especially useful in determining whether a policy is being enforced, in terms of performance. An increasing amount of traffic, however, does not rely on TCP and instead uses UDP. UDP does not contain the necessary information to determine service levels according to conventional approaches.

[0080] To determine service levels to these destinations, the present invention might employ a statistically relevant amount of collateral TCP traffic going to the same prefix or a series of active probes to the same destinations, or have the analyzer parse deeper into the packet and understand the traffic at the application layer (e.g., layer 7). There are some protocols running on UDP that have very specific requirements that are different from most other data traffic on the network. These protocols are loosely classified as "real-time" protocols and include things like streaming media and Voice over IP ("H.323"). Packet loss and latency, below a certain level, are secondary concerns for real-time protocols.

[0081] Most importantly, however, is reducing the variance in inter-packet arrival times (i.e., network jitter). Many real time protocols such as H.323 report the observed jitter in back channel communication known as the RTCP ("Real-Time Transport Control Protocol"), which is used to distribute time-dependent media data via IP multicast with feedback. If passive flow analyzer 630 of Figure 3 is "application-aware," it can capture and observe the contents of the RTCP and be aware when the underlying network path is not meeting minimum jitter requirements. This could trigger an SLA violation in the same manner that 30% packet loss would.

[0082] Correlator 652 operates to interpret and to group the packet elements (e.g., TCP and IP) from the packets to determine the current service level of the flow and then groups the packets into a specific traffic flow. Flows are reconstructed, or grouped, by matching source and destination IP addresses and port numbers, similar to process of stateful monitoring of firewalls. Correlator 252 determines the current service level by measuring several traffic characteristics during a TCP transaction. For example, correlator 252 determines the round trip time ("RTT") incurred on a network, and hence, this serves as a measure of latency for the network traffic.

[0083] Figure 7 shows how correlator 652 of passive flow analyzer 630 of Figure 6, placed near a source (e.g., client having a source address), can determine the network latency ("NL") and server response time ("SRT") for a TCP traffic stream. Figure 8 shows how correlator 652 of passive flow analyzer 630 of Figure 6, placed near a destination (e.g., server having a destination address), can determine the network latency ("NL") and server response time ("SRT") for a TCP traffic stream

[0084] Correlator 652 of Figure 6 determines NL, for example, by estimating the difference 791 of Figure 7 in time between a TCP SYN packet and its corresponding TCP SYN ACK packet. The difference in time between SYN and SYN ACK 791 is a rough estimation of the RTT excluding the small amount of time 790 that the server takes to respond to SYN. The SYN ACK packet is handled in the kernel of most operating systems and is generally assumed to be near zero. For each new TCP stream that is initiated from the source, correlator 652 can observe a time instantaneous value for network latency.

[0085] Packet loss is calculated, as a percentage, by correlator 652 by maintaining the state of all of the retransmitted packets that occur. From this value, correlator 652 calculates percentage packet loss from a total count of segments sent.

[0086] Correlator 652 also determines SRT 792 of Figure 7, for example, by estimating the delta time (i.e., difference) 793 between, for example, the HTTP GET message 795 and the first data segment received and then by subtracting the previous value for the RTT. This assumes that the previous value for the RTT has not changed beyond an operable range since the TCP handshake occurred. The measurement shown by 794 indicates that measured congestion increases in the path as SRT 792 correspondingly increases. For purposes of this example, it is assumed that the data segments in the initial HTTP GET are sent back to back. In Figure 7, the passive flow analyzer 630 is deployed close to (i.e., minimal or negligible latency due to geographically different locations) the clients requesting content from the IP data network, such as the Internet.

[0087] Correlator 652 also determines SRT 892 of Figure 8, for example, by estimating the delta time between the HTTP GET message 893 and the first data segment 894. In Figure 8, the passive flow analyzer 630 of Figure 6 is deployed on the server end as will occur for most content delivery installations.

[0088] Referring back to Figure 8, SRT 892 determined by correlator 652 depends on its location along the path that the traffic traverses. If passive flow analyzer 630 of Figure 6 is on the client side, server response time 792 of Figure 7 can be estimated as the delta in time between the HTTP GET Request message and the first data segment returned minus the RTT observed before the GET Request as shown in Figure 7. If passive flow analyzer 630 of Figure 6 is closer to the server side, the estimation is essentially the delta in time between the GET Request and the response as shown in Figure 8. Congestion estimations are also possible by using the TCP Congestion Window ("cwnd") and by identifying the delta in receive time between segments that were sent back to back by the server, where the TCP congestion window controls the number of packets a TCP flow may have in the network at any time. Correlator 652 is coupled to provide the above determined exemplary flow characteristics to aggregator 653.

[0089] Referring back to Figure 6, aggregator 653 primarily operates to group all flows going to each set of specific destinations together into one grouping. Aggregator 653 uses the service level statistics for each of the individual flows, received from Correlator 652, to generate an aggregate of service level statistics for each grouping of flows that are to go to the same destinations in the data network, such as the Internet. Aggregator 653 is also coupled to traffic storage 621 to store such aggregated (i.e., grouped by address prefix) traffic flow characteristics. Traffic flow characteristics (or traffic profiles) are then used for future statistical manipulation and flow prediction. In a specific embodiment, storage 621 is the equivalent, or the same, as storage 221 of Figure 2.

[0090] The granularity of the destinations is the same as the granularity of changes that can be made in the routing table. Nominally, flow control system of Figure 2 could install routes with prefixes of any length (i.e., 0/ to /32), though the general practice is not to do so. Aggregator 653, therefore, will start aggregating flow statistics at the /32 level (i.e., class C networks) and continue all the way up to the /8 level (i.e., class A networks) into a data structure, such as a patricia or radix trie, or a parent-child data structure, or the like. In this way, it is possible to seek very quickly the necessary granularity of the routing change that needs to be made to ensure the service level is met.

[0091] Aggregation techniques employed by aggregator 653 are used to maintain the system 200 of Figure 2 to acceptable performance service levels, such as determined by one or more flow policy requirements. Since network performance has been shown not to follow conventional statistical distribution, such as Gaussian or Poisson distribution, average calculations for service levels across all flows are not as reliable a measurement of a typical performance behavior during a pre-determined time interval. If the service level agreement (SLA) or policy, however, states that the average service level must be maintained, then the outlying occurrences of poor performance will cause the average to be skewed, thus requiring corrective action to restore the minimum service levels being offered. A meaningful way to describe typical service levels being offered across all flows is to use median values, rather than average values. A person having ordinary skill in the arts will appreciate that either technique is possible and will depend on the definition of the service level that must be maintained.

[0092] Figure 9 illustrates how passive flow analyzer 930, according to another embodiment of the present invention, is capable of packet capture and flow reconstruction across more than one network interface, each interface represented by a network interface card ("NIC"). In practice, many switch fabrics are constructed in a manner by tapping into a single point in the data stream or replicating a single port. The switch does not guarantee that passive flow analyzer 930 will see all of the traffic in both directions. Bi-directional traffic is required for optional flow reconstruction for passive analysis. In figure 9, the switch fabric shown must be passively tapped at tap points 921 at four places (as shown) and connected to passive flow analyzer 931 at four different network interface cards (NIC) 922. Passive taps at tap points 921 can be mirrored switch ports or optical/electrical passive taps. Passive flow analyzer 930 has a single or combined aggregated flow reconstruction element 953 that can collects captured data from multiple network interfaces in order to perform flow reconstruction.

[0093] Figure 10 illustrates yet another embodiment of the present invention where passive flow analyzer 630 of Figure 6 is distributed in nature. Figure 10 shows traffic flow 1020 bi-directionally traveling via several local traffic source points. Distributed local passive flow agents 1025 are tapped passively at tap point 1024 into traffic flow 1020. Passive flow agents 1025 are distributed such that each agent monitors and conveys individual flow characteristics. The traffic sources are distributed across a layer 3 infrastructure, for example, and are separated by one or more routers 1026. This arrangement prevents the passive flow analyzer 930 of Figure 9 from collecting information across the same layer 2 switch fabric as in Figure 9. Each of the passive flow agents 1025 performs local flow reconstruction and then exports flow data records 1027 over the network to a central passive flow analyzer 1028, performs flow aggregation and service level analysis across all of the distributed passive flow agents 1025.

[0094] Figure 11 illustrates a more detailed functional block diagram depicting multiple passive flow agents 1125 separately distributed and a single central passive flow analyzer 1128. Each passive flow agent 1125 includes packet capture 1150, parser 1151 and correlator 1152 functions on each of the local traffic flows. Correlator 1152 exports flow records 1129 with substantial data reduction to central passive flow analyzer 1128. Substantial data reduction is used to reduce the amount of information forwarded to the central passive flow analyzer and can be effectuated by using well-known encoding techniques. Central passive flow analyzer 1128 accepts flow export records 1129 from each passive flow agent 1125 and central aggregator 1153 performs prefix aggregation on each of the exported flows. Thus, the centrally aggregated flow information can be used to determine if a particular policy violation is occurring.

[0095] Figure 12 illustrates a detailed block diagram of usage collector 214 of Figure 2. Usage collector 1215 operates to collect usage information 1273 from network providers, such as byte counters (i.e., the amount of traffic transmitted to and received from network service providers). Usage collector 1215 uses this information to calculate network service provider utilization, load, etc., of data paths associated with the provider.

[0096] Usage collector 1215 also operates to reconstruct provider billing records. Usage collector 1215 accepts provider configuration information 1271 related to each network service provider (NSP) connection. This NSP configuration information 1271 details provider interfaces on the various routers 1272 (e.g., egress routers), provider next-hop IP addresses traceroute probes (to verify the current provider in use with trace probes), billing period start and end dates, circuit bandwidth for calculating the utilization and price per megabit/sec, minimum bandwidth commitment, burstable rates, provider sampling interval, provider billing algorithm, a utilization alarm threshold and the like.

[0097] In operation, exemplary raw collector 1274 sends a query 1290 (e.g., SNMP) to collect interface raw byte counters from routers 1272 on each of the provider circuits at a specified sampling interval. Provider circuits include paths, pipes virtual or physical, T1, and the like. Raw Collector 1274 places the raw byte counters 1280 into persistent storage for later reporting and analysis. Raw collector 1274 sends the raw information to two other components: utilization monitor 1275 and bill reconstructor 1276.

[0098] Utilization monitor 1275 calculates the ingress and egress circuit utilization for each provider using the raw byte counts and the NSP configuration information 1271. In one example, NSP configuration information 1271 includes the bandwidth of the provider's circuits. Utilization information 264 includes data representing utilization trends for use with short range forecasting models (e.g., ARIMA, exponential smoothing, etc.) such that utilization monitor 1275 can determine whether bandwidth is trending up or down (i.e., increasing or decreasing in size) for a given service provider.

[0099] Bill reconstructor 1276 uses the billing information from NSP configuration data 1271 to reconstruct the current provider billable rate for the current billing period. Billing information includes information explaining the methods that specific providers use to calculate costs, such as a billing rate. Such methods of calculating bills for using a network provider are well known in the art. Bill reconstructor 1276 applies similar provider billing methods to the raw byte counters from raw collector 1274 to generate the bill and related billing rates, etc. The generated bills, which are mapped into dollar amounts, are typically estimates since the sample times between the provider and usage collector 1215 will not match exactly. Bill reconstructor 1276 will send billing information 1261 to controller 1202 for use in peak avoidance and least cost routing. Peak avoidance is defined as a method of avoiding using a path or path segment at a higher a billing rate, such as shown in Figure

15. Least cost routing refers to a method of using or defaulting traffic to the least expensive provider.

[00100] Additionally the information can be sent to controller 1202 for use in the least cost fix method of selecting the cheapest if performance is of no consequence. That is, controller 1202 uses data from billing message 1261, including billing rates, to determine an alternate route based in part on a route's free bandwidth (i.e., route does not incur additional cost to use), in accordance with the flow policy.

[00101] Referring back to Figure 2, configuration element 211 is coupled to controller 205 and to data director 220. Controller 205 provides the best route to reach a destination prefix to configuration element 211. Configuration element 211 operates to change the default routing behavior (i.e., current path) for the destination requiring corrective action. Configuration element 211 changes the routing behavior by, for example, sending a modified routing table of addresses to data director 220.

[00102] Once data director 220 receives this information, direct 220 informs controller 205 that route change has been implemented. Thereafter, controller 205 communicates signal 230 back to passive calibrator 202 to clear its state and to resume monitoring the destination. The destination is monitored to ensure that the updated route of the routing table, or path, meets minimum service levels (e.g., no violations of SLA, or no unacceptable deviations from agreed upon performance metrics as defined by the associated flow policy).

[00103] In one aspect, configuration element 211 resides in a route server. In another aspect, configuration element 211 resides in a router and is configured to modify a route map or table. In yet another aspect, configuration element 211 is adapted to provide configuration information, or routing table. In still yet another aspect, the route information is stored within

the configuration element 211 according to whether it is related to inbound or outbound traffic.

[00104] Figure 13 shows an example of yet another embodiment of the present invention, where configuration element 211 of Figure 2 resides in a network element, such as route server 1391. Configuration element 1384 of Figure 13 operates similarly to other adaptations of configuration elements described herein. That is, configuration element 1384 modulates the current or default routes of data traffic and thus modifies the default routing behavior, for example, in a local deployment (e.g., Point of Presence, or "POP"). Route server 1391 ("RS") receives a full set or sub-set of routing tables from the data networks of interest.

[00105] In one embodiment, the routing tables are received into route server 1391 by way of one or more default BGP4 feeds 1392 into BGP4 Engine 1382 from a full set or sub-set of the local transit providers. BGP4 Engine 1382 integrates, or merges, all of the routes into a single BGP4 routing table 1383 best available routes. In another embodiment, route server 1391 maintains an iBGP session with all of the internal BGP capable routers rather than maintaining the BGP4 sessions as shown in Figure 13. With a single iBGP session there is no need to configure all of the BGP sessions with the network service providers before making route changes.

[00106] Configuration element 1384 is designed to receive one or more BGP4 routing tables 1383 from BGP4 engine 1382 and is adapted to receive one or more control signals and data resulting from the control processes of controller 1305. In operations, configuration element 1384 receives, from controller 1305, the necessary routing changes to be implemented in default routing table 1388. Then, configuration element 1384 incorporates one or more changes in modified routing table 1389.

[00107] Thus, configuration element 1384 operates to modify BGP4 routing table 1383 and to generate one or more modified BGP4 routing tables 1388. Modified BGP4 routing table 1388 includes changed routing 1389, advertisements of more specific routes, etc. New modified BGP4 routing table 1388 is then fed to all BGP clients in the network, which then is used to guide traffic to the destination.

[00108] For a given source address, the ingress point into a network is determined typically by the advertisements of routes made to downstream providers and a provider policy (set of rules that is set up by such providers). Eventually, the network service provider (e.g., "ISP") that is hosting the destination will receive such advertisements.

[00109] Controller 205 of Figure 2 is designed to receive performance characteristics, such as latency, loss, jitter, etc., as monitored by the calibrator elements as well as usage characteristics, such as bandwidth, costs, etc., as monitored by the usage collector. Controller 205 is coupled to policy repository 218 to receive flow policies, which typically include service level agreement ("SLA") performance metrics. These metrics, or requirements, are compared against the monitored performance and usage characteristics . If a particular policy is violated (i.e., one or more performance metrics are outside one or more expected ranges or values), controller 205 determines a sub-set of one or more alternate data paths that conform to the associated flow policy. In another example, controller 205 selects a best or optimized path as an alternate data path that best meets the performance requirements and usage requirements, as defined by the policy.

[00110] The active calibrator and the passive calibrator provide performance characteristics. Regarding the active calibrator, controller 205 initiates active calibration by request active probing. The active calibrator sends one or more calibration probes on probe path 207 out into the one or more data networks. The returning probes on probe path 207

provide information back to controller 205, which contains the identities of available paths and performance information related thereto.

[00111] Regarding the passive calibrator, controller 205 is designed to receive real- or near-real time network performance characteristics (i.e., loss, latency, jitter, etc.) from passive calibrator 230 as monitor in traffic flows in which it has access. After, controller 205 provides a routing change, or update, to configuration element 211, it also communicates a signal 230 to passive calibrator 203 when an updated route change is made to a specific destination. Signal 230 initiates the clearing of the state of passive calibrator 203 so that the calibrator resumes monitoring the specific destination to ensure that the updated route of the routing table, or path, is flow policy compliant. Clear state signal 338 of Figure 3 depicts the signal that comes from the controller to initiate the resetting of the passive flow analyzer's state.

[00112] In one example, controller 205 operates to interpret the aggregated flow data over an interval of time for each of the groupings of destination prefixes. And if a policy violation occurs, controller 205 determines which of the alternate routes, or paths, are best suited for the prefix or traffic type associated with the current traffic flow. Controller 205 then sends the necessary routing changes to configuration element 211. That is, controller 205 resolve policy violations relating to non-compliant network performance characteristics, in accordance with the associated flow policy. This process is repeated until the policy violation is resolved.

[00113] In another example, controller 1202 of Figure 12 is designed to receive real- or near-real time data representing network usage characteristics from usage collector 1215, such as usage rate, billing rates, etc. Controller 1202 uses this information to resolve policy violations relating to non-compliant usages characteristics, in accordance with the associated flow policy. That is, prior to or during a route change, controller 1202 not only does the

controller consider the performance of alternate paths, but also whether those alternate paths either avoid peak data traffic over a specific provider's path (i.e., adequate bandwidth related to turn-of-day) or are the least cost paths in view of the flow policies.

[00114] To resolve usage-type policy violations, controller 1202 is configured to receive routing tables, for example, to determine which of the current traffic flows or routing of data on certain paths, or path segments thereof, are congested (i.e., loaded) with respect to a particular provider path or paths. Controller 1202 also is designed to receive data representing flow volumes for each of the alternate provider paths to determine which sub-set of flows of a set of traffic flows to or from a given destination prefix are in compliance with the associated flow policy in terms of traffic flow volume.

[00115] An exemplary controller of the present thus is designed to obtain information related to the performance and usage of data networks and the make corrective action to effectively and efficiently route data over paths or segment of paths that meet at least associated policy requirements.

[00116] The following discussion relates to flow policies and the application of such policies in resolving policy violations and in enforcing the policy requirements or metrics. Referring back to Figure 2, controller 205 is coupled to policy repository 218 for receiving one or more policies. As described above, a policy is a set of rules or threshold values (i.e., maximums, minimums, and ranges of acceptable operations) that controller 205 uses these rules to compare against the actual flow characteristics of a specific traffic flow. For example, a policy is the user-defined mechanism that is employed by controller 205 to detect specific traffic flows that are to be monitored and acted upon if necessary. As an example, a policy can also specify how the particular policy should be enforced (i.e., in includes a hierarchical structure to resolve violations from highest to lowest precedence). Although an exemplary policy includes requirements, or rules, related to detection, performance, cost, and

precedence, one having ordinary skill the art should appreciate that less, or additional parameters, can be measured and enforced according the present invention.

[00117] Detection is defined as the techniques or mechanisms by which flow control system 200 determines which traffic that should be acted upon in response to a policy violation. The traffic flow can be identified, by name, by source or destination addresses, by source or destination ports, or any other known identification techniques. For example, a policy can be associated to only prefix. That is, system 200 will monitor the traffic flow to and from a specific prefix, and if necessary, will enforce the associated flow policy in accordance to its requirements. Further regarding detection, a policy defined for more specific prefixes can take precedence over more general prefixes. For example, a policy defined for a /24 will take precedence over a /16 even if the /16 contains the specific /24.

[00118] Performance is a policy requirement that describes one or more target performance levels (i.e., network/QoS policy parameters) or thresholds applied to a given prefix or prefix list. Although more than one performance-based policy requirement may be defined, in this example only a single policy is applied to a given prefix or prefix list. Exemplary performance requirements include loss, latency, and jitter.

[00119] Moreover, such requirements can be configured either as, for example, an absolute, fixed value or as an Exponentially Weighted Moving Average ("EWMA"). Absolute value establishes a numerical threshold, such as expressed as a percentage or in time units over a configurable time window. The EWMA method establishes a moving threshold based on historic sampling that places an exponential weighting on the most recent samples, thereby asserting a threshold that can take into account current network conditions as they relate to historic conditions.

[00120] Cost is expressed in the policy definition in terms of precedence and whether the policy is predictive or reactive. Costs are characterized by usage collector 214 of figure 2 through bill reconstruction and reconciliation of bandwidth utilization in both aggregate and very granular levels (e.g., by /24 destination network). Cost predictive requirements are used to proactively divert traffic from one provider to another in order to avoid establishing a peak (i.e., "peak avoidance") that may trigger a new or higher billable rate. Cost reactive requirements are used to reactively divert traffic from one provider to another when a minimum commit rate or current billable rate is exceeded.

[00121] . Typically, both cost predictive and reactive requirements result in a binary decision (i.e., a circuit or path, for example, is either in compliance with or in violation of a flow policy). In the case of predictive cost, the transit circuit is either in compliance, or soon to be violation of a flow policy. Regardless, an action must be taken to resolve the situation, unless cost is preceded by performance (i.e., performance requirements are to be addressed prior to making a cost-based change).

[00122] Precedence is a policy requirement that describes one or more target usage or utilization characteristics or levels. Precedence includes provider preference and maximum utilization (i.e., load) requirements. The provider preference requirement is, for example, an arbitrary ranking of providers that is used when an action must be taken, but when two or more transits may be selected in order to enforce the policy. The flow control system can automatically set the provider or path preference requirement if it is not configured explicitly by the system's operator. This requirement is then applied as a tiebreaker in deadlocked situations such that the provider with the highest preference wins the tie and thus receive the diverted traffic flow.

[00123] The maximum usage requirement can be used as either may also be used an actual operational threshold not to be exceeded or as a tiebreaker. Maximum usage is configured, for example, in the transit provider section of the configuration and takes either a percentage argument (i.e., in terms of available bandwidth), or alternatively, can be set as an absolute value in terms of Mb/s (i.e., not to exceed available bandwidth).

[00124] The following is an example of a policy used with a controller to determine whether the specific policy is in compliance, and if not, to determine the course of action.

[00125] For example, consider the following policy is used for a particular traffic flow:

Policy Requirement	Precedence	Value or Threshold
Loss	10	2%
Latency	20	EWMA
Cost	30	Predictive
Maximum usage	40	
Provider Preference	50	

[00126] Suppose that traffic flow is associated with prefix 24.0.34.0/24, is currently carrying traffic at 240 kbits/sec, and is reached via provider 1 of 3. Provider 1 is currently carrying 2 Mbits/sec and has a minimum commit of 5 Mbits/sec.

[00127] The controller of the flow control system using the policy can monitor the alternate traffic routes, or paths, and can determine the following flow characteristics as they relate to the providers:

Requirement	Value for ISP1	Value for ISP2	Value for ISP3
Loss	5% (violation)	Not available	Not available
Latency	140 ms	Not available	Not available
Cost	In compliance	In violation	In violation
Max Usage/	5 Mb/s	5 Mb/s	5 Mb/s
as Measured	2 Mb/s (compliance)	4 Mb/s (compliance)	5.5 Mb/s (violation)
Latency	100ms	100ms	100ms

[00128] In this case, ISP1 is in a violation state since loss of 5% exceeds the maximum loss requirement of 2% and since loss has been designated with the precedence of 10, with 50 being the lowest. Corrective action must be taken. The policy will be enforced without latency or loss information (i.e., because there is, for example, no visibility into the performance of the other links). In this case, the controller may initiate active probing using

the active calibrator to determine whether the other ISPs (including ISP2 and ISP3) are in compliance. Alternatively, the controller might determine the course of action based on the next parameter in the policy where the requirement is known (e.g., cost in this case). Since ISP 2 is in compliance and ISP 3 is not, ISP 2 would be chosen by the controller. If the two were both in compliance, the controller would go to the next ranked requirement, which is MaxUtil. If this is the case, ISP2 would is still selected.

[00129] In summary, the policy, such as the above exemplary policy, is input into the controller 205 of Figure 2 and is associated with, for example, a specific prefix. The general detection method (absolute or baseline/historical) can be specified as per prefix, thus specifying hard or absolute thresholds for some destinations that are well known, while using a baseline method for other destinations. The policy also defines the resolution method (e.g. procedure) to be used in the combination with performance metrics that must be met before the violation is considered resolved. Other parameters such as cost and utilization thresholds can be set per prefix. This gives the controller an indication of which prefixes should never be moved for cost or utilization reasons and which prefixes should be moved under any circumstances.

[00130] In order for controller 205 to handle peering connections, controller 205 communicates with the data director 220 to retrieve reachability information (i.e., routing tables) for the specific prefix that is about to be changed. In the case of transit circuits, controller 205 uses active calibrator 207 to determine reachability information (i.e., routing tables) for a given destination by, for example, sending active probes to the destination and then waiting for the response. Although peering connections are often unreachable, it is possible for active probes to succeed since some providers may not effectively filter traffic at a peering point and instead rely on an honor-like system to ensure that only traffic to those advertised destinations is received.

[00131] Therefore, in the case of peering, controller 205 must look in the routing table for an advertisement of that destination before moving traffic to a peering connection. Referring to Figure 15, iBGP feed 1599 includes advertised inactive routes as well as active routes. Otherwise, data director 220 of Figure 2 can be configured in accordance to route server 1591 of Figure 13, where eBGP is available from all providers.

[00132] Figure 14 illustrates how the availability of "free" bandwidth is expressed for a given provider and as measured by usage collector 214 of Figure 2. Over any given time period from t0 though t1, current usage rate 1602 and the current billable rate 1600 is determined. As shown, time point t0.5 1603 represents an over-sampled time point. Difference 1601 between these two values represents an amount of bandwidth available to be used without incurring any additional cost. The free bandwidth per provider can be used to select a sub-set of compliant providers when a performance-based policy is in violation by the current or default provider. Additionally, this information is used to apply cost- and load-based policies for each provider.

[00133] Figure 15 depicts how usage collector 214 calculates the time-continuous billable rate as shown in Figure 14. Most providers start out with a minimum commitment level 1710. If the current usage starts out below that commitment, the free bandwidth 1711 is shown. Samples are collected at twice the provider sampling rate to ensure that an accurate rate is being calculated (i.e., this is a conservative estimate and if the rate deviates from the provider rate, it will be higher and represent an overestimation of the billable rate). The small tick marks on the time axis represent the samples collected by the system (i.e., over-sampling). When enough samples are collected, the billable rate, which generally is expressed as the 95th percentile of all rate samples, may exceed the minimum commitment as shown by successively higher tiers 1713 of the billable rate in Figure 15. When the traffic drops back down below this rate, a new billable rate 1714 is set and the system again has free bandwidth 1718 available for use.

[00134] Figure 16 shows how an exemplary system 200 will detect a cost-based policy violation. Suppose the cost policy requirement is defined to be an absolute threshold, as shown by 1813. This threshold can be an absolute rate or a set dollar amount to spend (which is converted by the system to an average billable rate). On a sample-by-sample basis, the actual traffic rate 1814 should be such that a new billable rate above 1813 is never established. Using short range forecasting techniques, the traffic rate for the next few samples 1815 can be forecasted, and if this forecast predicts that a new billable rate 1816 will be established, controller 205 of Figure 2 can react by moving traffic off of this provider.

[00135] Although the present invention has been discussed with respect to specific embodiments, one of ordinary skill in the art will realize that these embodiments are merely illustrative, and not restrictive, of the invention. For example, although the above description describes the network communication data as Internet traffic, it should be understood that the present invention relates to networks in general and need not be restricted to Internet data. The scope of the invention is to be determined solely by the appended claims.

[00136] In the foregoing specification, the invention is described with reference to specific embodiments thereof, but those skilled in the art will recognize that while the invention is not limited thereto. Various features and aspects of the above-described invention may be used individually or jointly. Further, although the invention has been described in the context of its implementation in a particular environment and for particular applications, its usefulness is not limited thereto and it can be utilized in any number of environments and applications without departing from the broader spirit and scope thereof. The specification and drawings are, accordingly, to be regarded as illustrative rather than restrictive.

WHAT IS CLAIMED IS:

1. A method of enforcing a policy for data communicated by a computer network designed to route data between a first point and a second point, the first point is coupled to one or more first networks, at least one of the one or more first networks is coupled to at least one of a plurality of second networks, at least one of the second networks is coupled to the second point, each of the networks includes at least one segment of a path, the path is from the first point to the second point, for transporting the data communicated to the second point, where at least two of the networks are coupled at an interconnection and where the data communicated flows through the interconnection point, the method comprising:

monitoring at least one usage characteristic associated with at least one segment, the at least one segment being located in a first network; and

comparing the at least one usage characteristic with an associated usage requirement of a policy.

2. The method of claim 1, further comprising:

determining if the at least one usage characteristic associated with the routing of data in the first network violates the usage requirement.

3. The method of claim 2, further comprising:

modifying the routing of data such that the at least one usage characteristic associated with the routing of data in the first network no longer violates the usage requirement.

4. The method of claim 3, wherein modifying the routing of data comprises:

monitoring at least one usage characteristic associated with at least another segment, the another segment being located in another first network;

determining that one or more usage characteristic associated with the routing of data in the another first network complies with the usage requirement; and

altering the routing of data such that data is routed thorough the another segment located in the another first network.

5. The method of claim 1, further comprising:

monitoring at least one performance characteristic associated with the path.

6. The method of claim 5, further comprising:

determining if the at least one performance characteristic associated with the path violates one or more performance requirements;

monitoring at least one performance characteristic associated with another path;

determining that one or more performance characteristics associated with the routing of data in the another path complies with the performance requirement; and

altering the routing of data such that data is routed through the another path.

7. The method of claim 1, wherein the monitoring at least one usage characteristic comprises:

measuring a usage characteristic related to utilization of the at least one segment.

8. The method of claim 1, wherein the monitoring at least one usage characteristic comprises:

measuring a usage characteristic related to usage of the at least one segment.

9. The method of claim 1, wherein the monitoring at least one usage characteristic comprises:

measuring a usage characteristic related to load of the at least one segment.

10. The method of claim 1, wherein the monitoring at least one usage characteristic comprises:

measuring a usage characteristic related to cost of the at least one segment.

11. The method of claim 1, wherein monitoring at least one performance characteristic comprises:

measuring a performance characteristic related to loss.

12. The method of claim 1, wherein monitoring at least one performance characteristic comprises:

measuring a performance characteristic related to latency.

13. The method of claim 1, wherein monitoring at least one performance characteristic comprises:

measuring a performance characteristic related to jitter.

14. A system for enforcing a policy for data communicated by a computer network designed to route data between a first point and a second point, the first point is coupled to one or more first networks, at least one of the one or more first networks is coupled to at least one of a plurality of second networks, at least one of the second networks is coupled to the second point, each of the networks includes at least one segment of a path, the path is from the first point to the second point, for transporting the data communicated to the second point, where at least two of the networks are coupled at an interconnection and where the data communicated flows through the interconnection point, the method comprising:

a usage monitoring module configurable to monitor at least one usage characteristic associated with at least one segment, the at least one segment being located in a first network; and

a comparing module configurable to compare the at least one usage characteristic with an associated usage requirement of a policy.

15. The system of claim 14, further comprising:

a usage determining module configurable to determine if the at least one usage characteristic associated with the routing of data in the first network violates the usage requirement.

16. The system of claim 15, further comprising:

a modifying module configurable to modify the routing of data such that the at least one usage characteristic associated with the routing of data in the first network no longer violates the usage requirement.

17. The system of claim 16, wherein the modifying module comprises:

a first module configurable to monitor at least one usage characteristic associated with at least another segment, the another segment being located in another first network;

a second module configurable to determine that one or more usage characteristic associated with the routing of data in the another first network complies with the usage requirement; and a third module configurable to alter the routing of data such that data is routed thorough the another segment located in the another first network.

18. The system of claim 14, further comprising:

a performance monitoring module configurable to monitor at least one performance characteristic associated with the path.

19. The system of claim 18, further comprising:

a performance determining module configurable to determine if the at least one performance characteristic associated with the path violates one or more performance requirements,

the performance monitoring module is configurable to monitor at least one performance characteristic associated with another path;

a performance determining module configurable to determine that one or more performance characteristics associated with the routing of data in the another path complies with the performance requirement; and

an altering module configurable to alter the routing of data such that data is routed through the another path.

20. The system of claim 14, wherein the usage monitoring module comprises:

a utilization measuring module configurable to measure a usage characteristic related to utilization of the at least one segment.

21. The system of claim 14, wherein the usage monitoring module comprises:

a usage measuring module configurable to measure a usage characteristic related to usage of the at least one segment.

22. The system of claim 14, wherein the usage monitoring module comprises:

a load measuring module configurable to measure a usage characteristic related to load of the at least one segment.

23. The system of claim 14, wherein the usage monitoring module comprises:

a cost measuring module configurable to measure a usage characteristic related to cost of the at least one segment.

24. The system of claim 14, wherein the usage monitoring module comprises:

a loss measuring module configurable to measure a performance characteristic related to loss.

25. The system of claim 14, wherein the usage monitoring module comprises:

a latency measuring module configurable to measure a performance characteristic related to loss.

26. The system of claim 14, wherein the performance monitoring module comprises:

a jitter measuring module configurable to measure a performance characteristic related to jitter.

27. A system for enforcing a policy for data communicated by a computer network designed to route data between a first point and a second point, the first point is coupled to one or more first networks, at least one of the one or more first networks is coupled to at least one of a plurality of second networks, at least one of the second networks is coupled to the second point, each of the networks includes at least one segment of a path, the path is from the first point to the second point, for transporting the data communicated to the second point, where at least two of the networks are coupled at an interconnection and where the data communicated flows through the interconnection point, the method comprising:

means for monitoring at least one usage characteristic associated with at least one segment, the at least one segment being located in a first network; and

means for comparing the at least one usage characteristic with an associated usage requirement of a policy.

28. The system of claim 27, further comprising:

means for determining if the at least one usage characteristic associated with the routing of data in the first network violates the usage requirement.

29. The system of claim 28, further comprising

means for modifying the routing of data such that the at least one usage characteristic associated with the routing of data in the first network no longer violates the usage requirement.

30. A computer-readable media for enforcing a policy for data communicated by a computer network designed to route data between a first point and a second point, the first point is coupled to one or more first networks, at least one of the one or more first networks is coupled to at least one of a plurality of second networks, at least one of the second networks is coupled to the second point, each of the networks includes at least one segment of a path, the path is from the first point to the second point, for transporting the data communicated to the second point, where at least two of the networks are coupled at an interconnection and where the data communicated flows through the interconnection point, the method comprising:

instructions for monitoring at least one usage characteristic associated with at least one segment, the at least one segment being located in a first network; and

instructions for comparing the at least one usage characteristic with an associated usage requirement of a policy.

31. The computer-readable media of claim 30, further comprising:

instructions for determining if the at least one usage characteristic associated with the routing of data in the first network violates the usage requirement.

32. The computer-readable media of claim 31, further comprising:

instructions for modifying the routing of data such that the at least one usage characteristic associated with the routing of data in the first network no longer violates the usage requirement.

33. A method of enforcing a policy for data communicated by a computer network designed to route data between a first point and a second point, the first point is coupled to one or more first networks, at least one of the one or more first networks is coupled to at least one of a plurality of second networks, at least one of the second networks is coupled to the second point, each of the networks includes at least one segment of a path, the path is from the first point to the second point, for transporting the data communicated to the second point, where at least two of the networks are coupled at an interconnection and

where the data communicated flows through the interconnection point, the method comprising:

comparing the at least one usage characteristic with an associated usage requirement of a policy;

determining if the at least one usage characteristic associated with the routing of data in the first network violates the usage requirement; and

modifying the routing of data such that the at least one usage characteristic associated with the routing of data in the first network no longer violates the usage requirement.

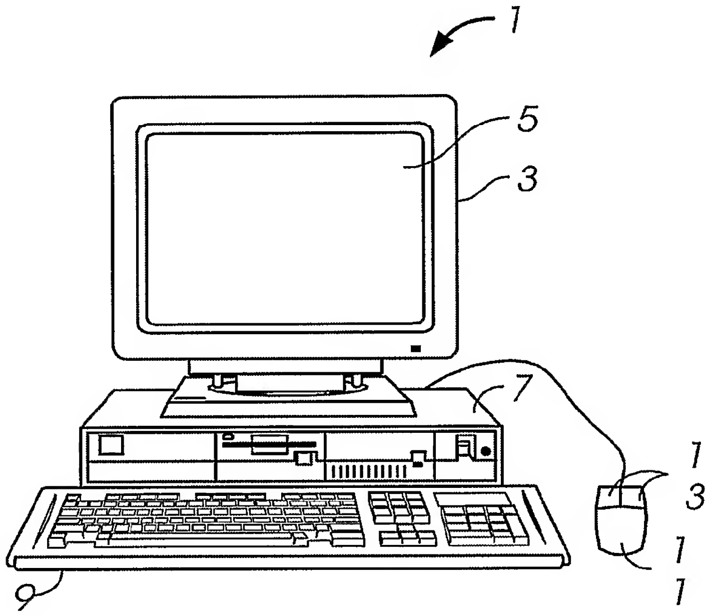

1
5
3
7
1
3
1
1
9

FIG. 1A

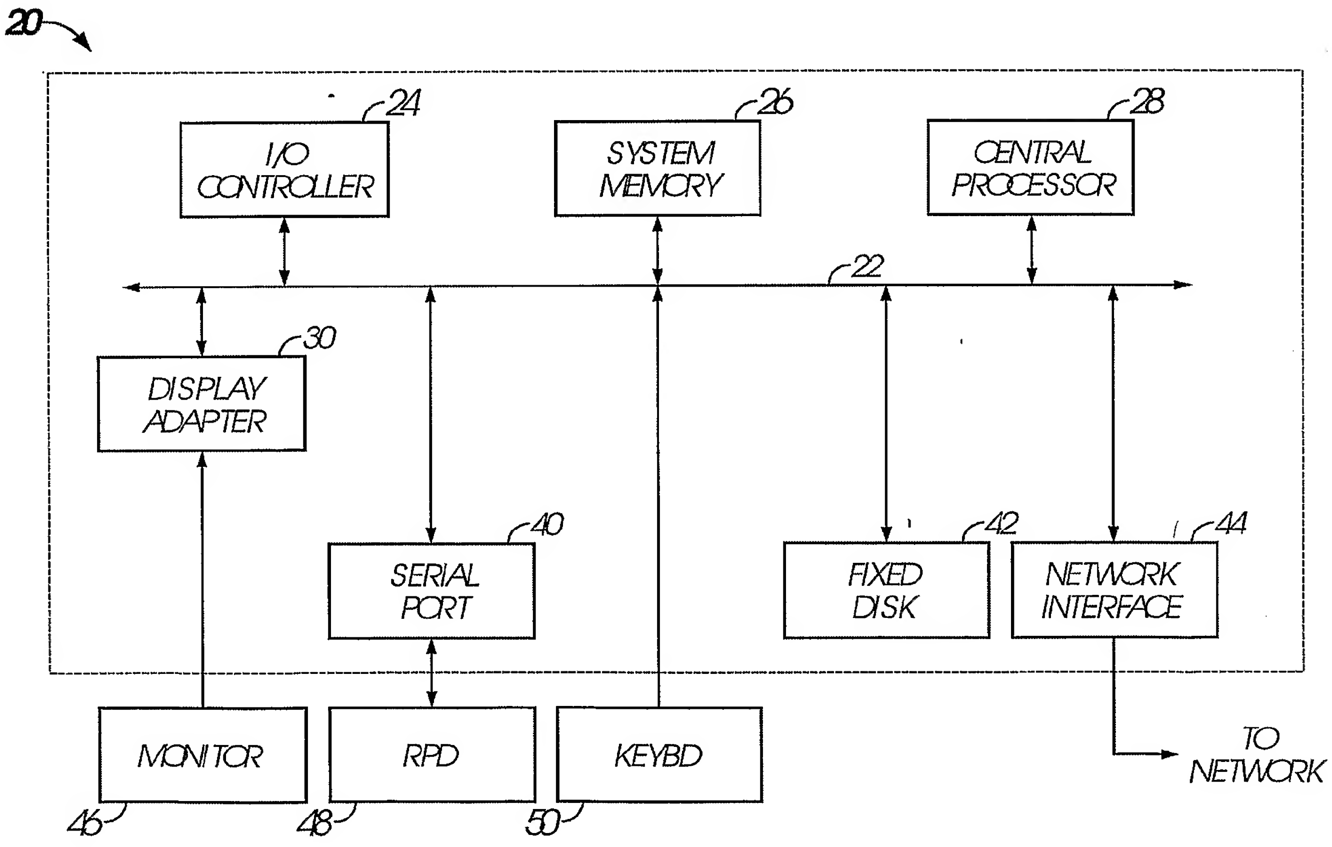

20
24
I/O CONTROLLER
26
SYSTEM MEMORY
28
CENTRAL PROCESSOR
22
30
DISPLAY ADAPTER
40
SERIAL PORT
42
FIXED DISK
44
NETWORK INTERFACE
MONITOR
RPD
KEYBD
TO NETWORK
46
48
50

FIG.1B

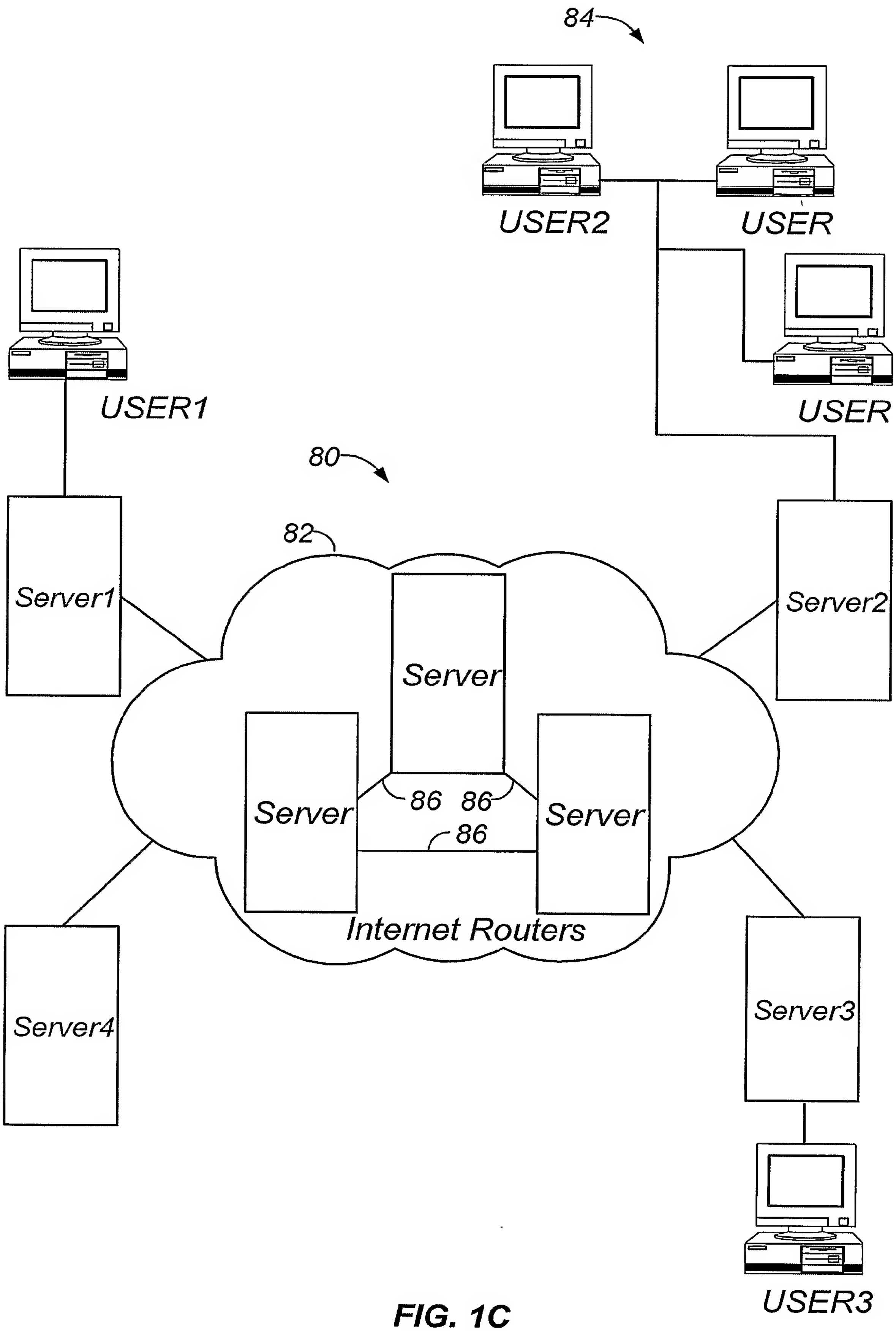

84
USER2
USER
USER1
USER
80
82
Server1
Server2
Server
Server
Server
86
86
86
Internet Routers
Server4
Server3
USER3

FIG. 1C

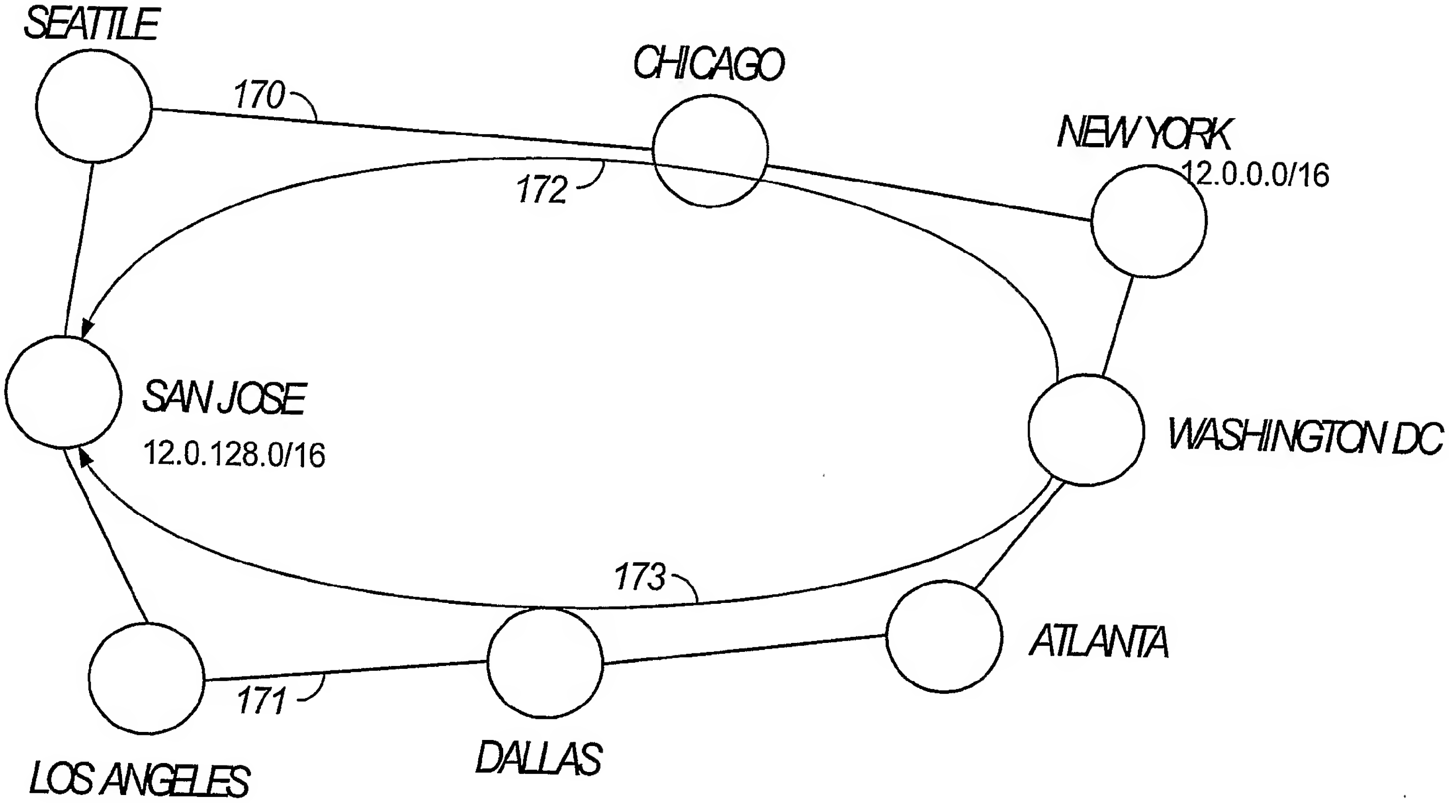

SEATTLE
CHICAGO
NEW YORK
12.0.0.0/16
170
172
SAN JOSE
12.0.128.0/16
WASHINGTON DC
173
ATLANTA
171
LOS ANGELES
DALLAS

FIG. 1D

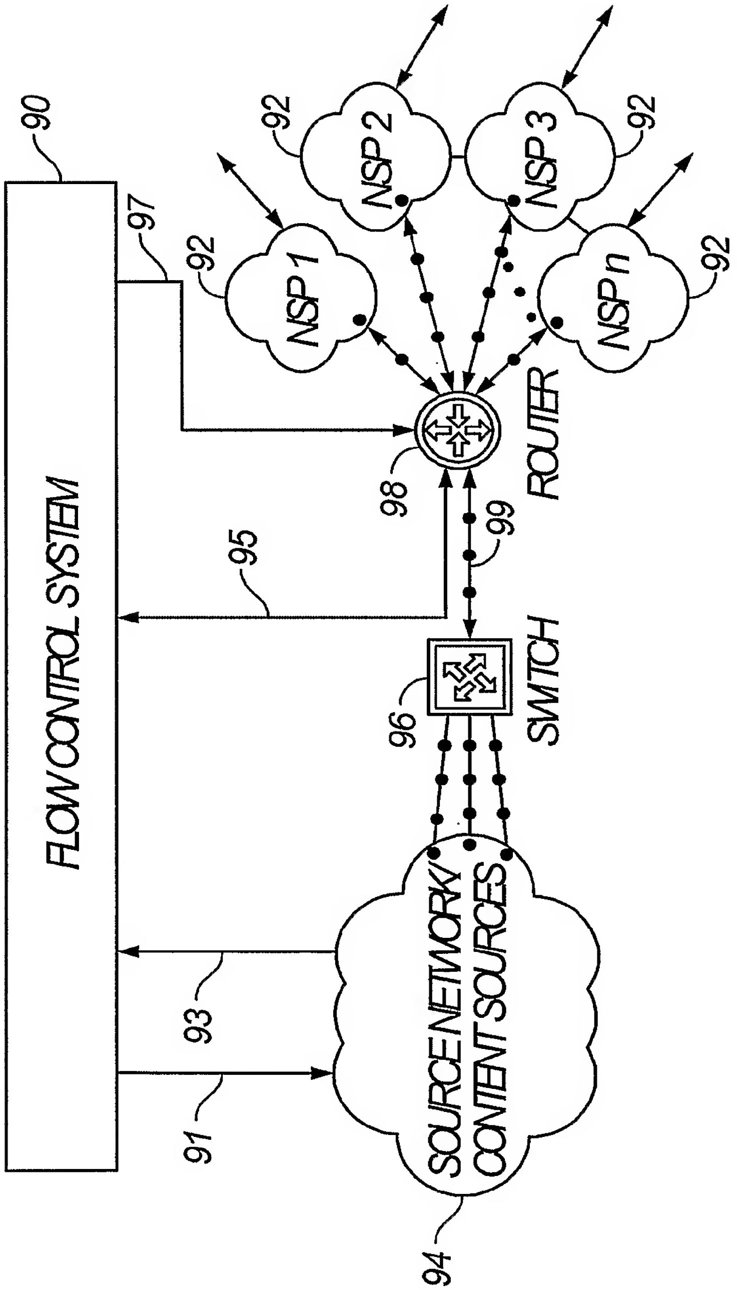

90
FLOW CONTROL SYSTEM
97
95
93
91
94
SOURCE NETWORK/
CONTENT SOURCES
96
SWITCH
99
98
ROUTER
92
NSP 1
92
NSP 2
92
NSP 3
92
NSP n

FIG 1E

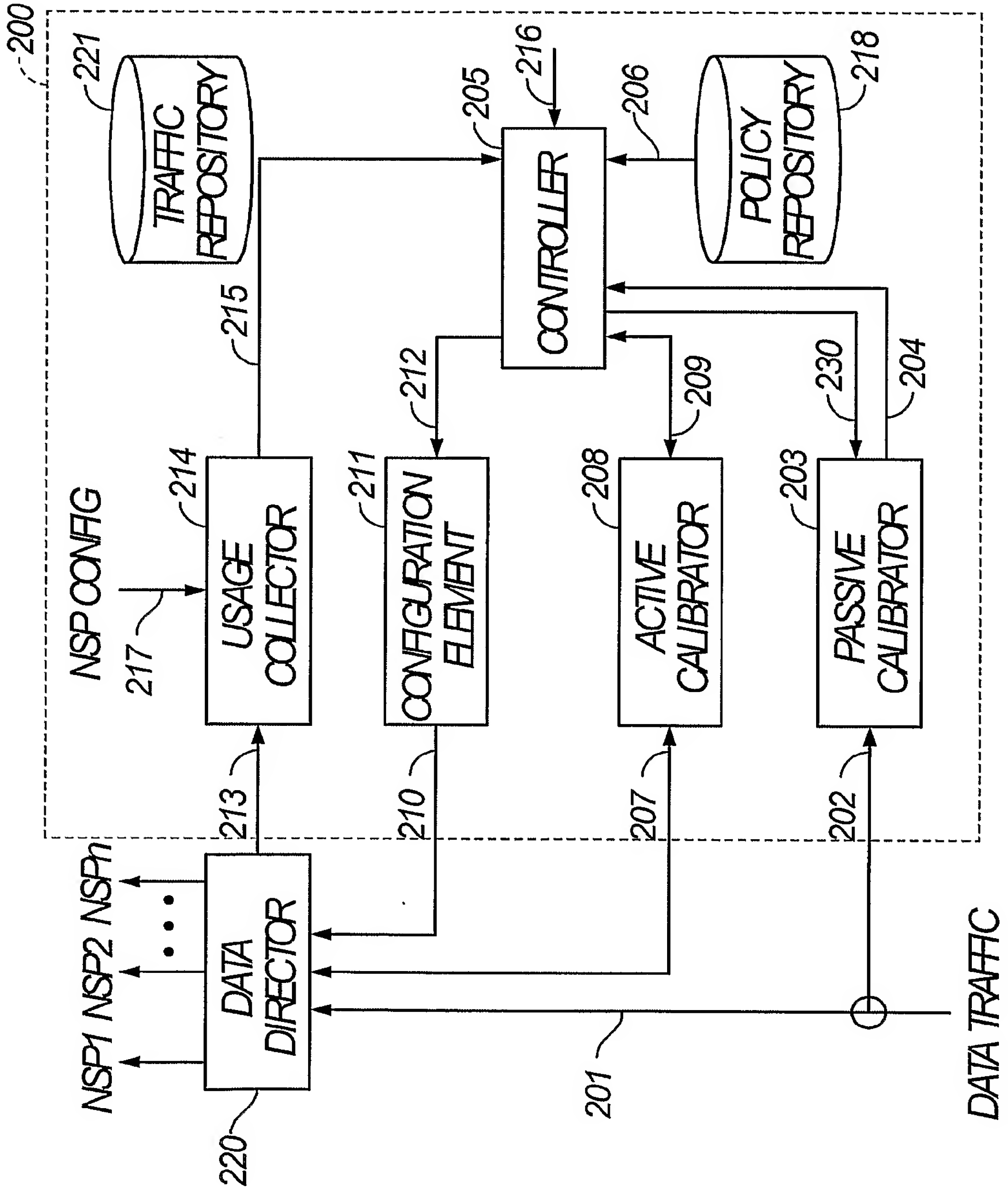
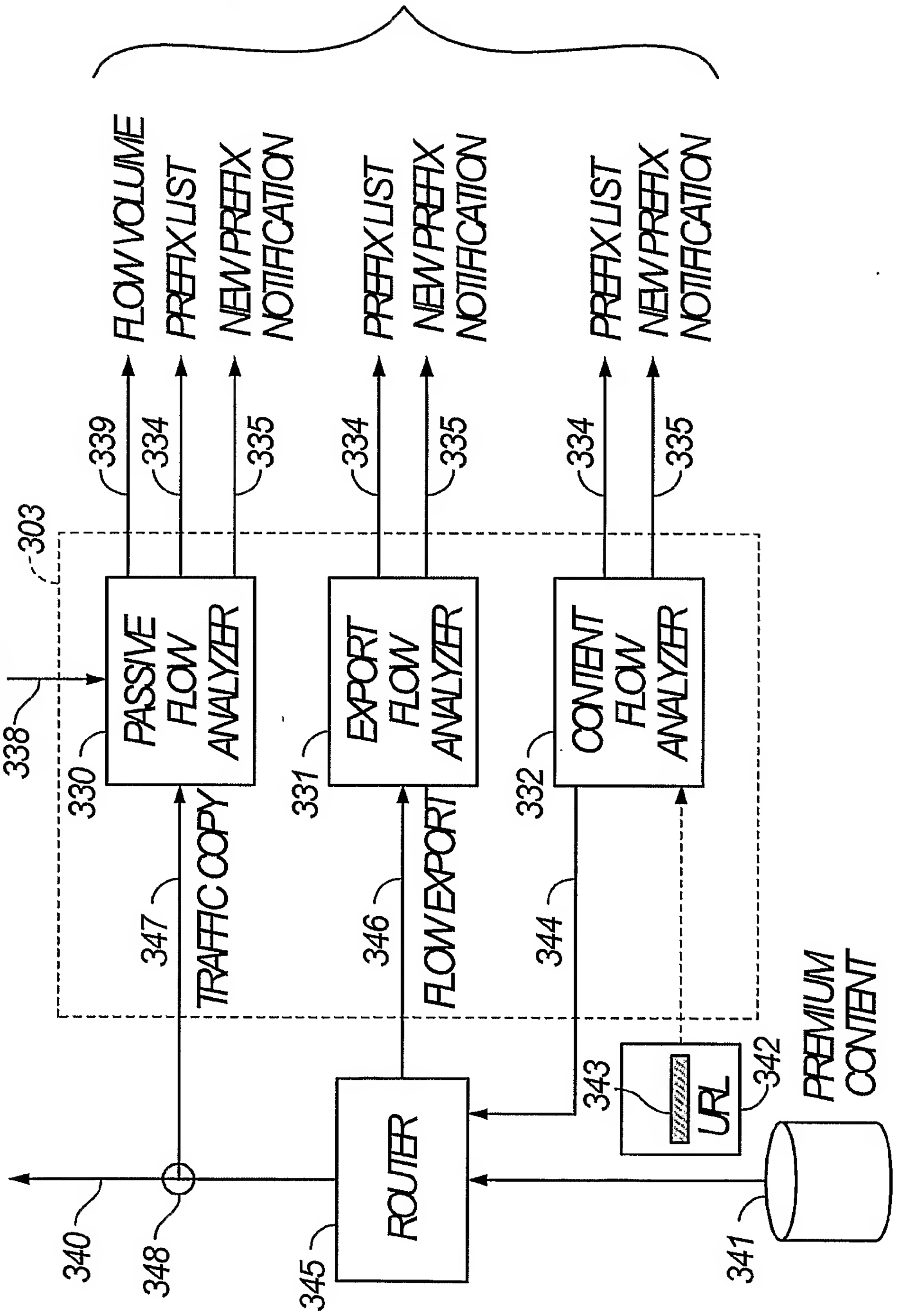

200
221
TRAFFIC REPOSITORY
205
216
206
CONTROLLER
POLICY REPOSITORY
218
215
212
209
230
204
NSP CONFIG
217
214
USAGE COLLECTOR
211
CONFIGURATION ELEMENT
208
ACTIVE CALIBRATOR
203
PASSIVE CALIBRATOR
213
210
207
202
NSP1 NSP2 NSPn
DATA DIRECTOR
220
201
DATA TRAFFIC

FIG 2



PREMIUM CONTENT
341
URL
342
343
ROUTER
345
340
348
344
346
FLOW EXPORT
347
TRAFFIC COPY
303
CLEAR STATE
338
PASSIVE FLOW ANALYZER
330
EXPORT FLOW ANALYZER
331
CONTENT FLOW ANALYZER
332
339
334
335
334
335
334
335
FLOW VOLUME
PREFIX LIST
NEW PREFIX NOTIFICATION
PREFIX LIST
NEW PREFIX NOTIFICATION
PREFIX LIST
NEW PREFIX NOTIFICATION

FIG 3

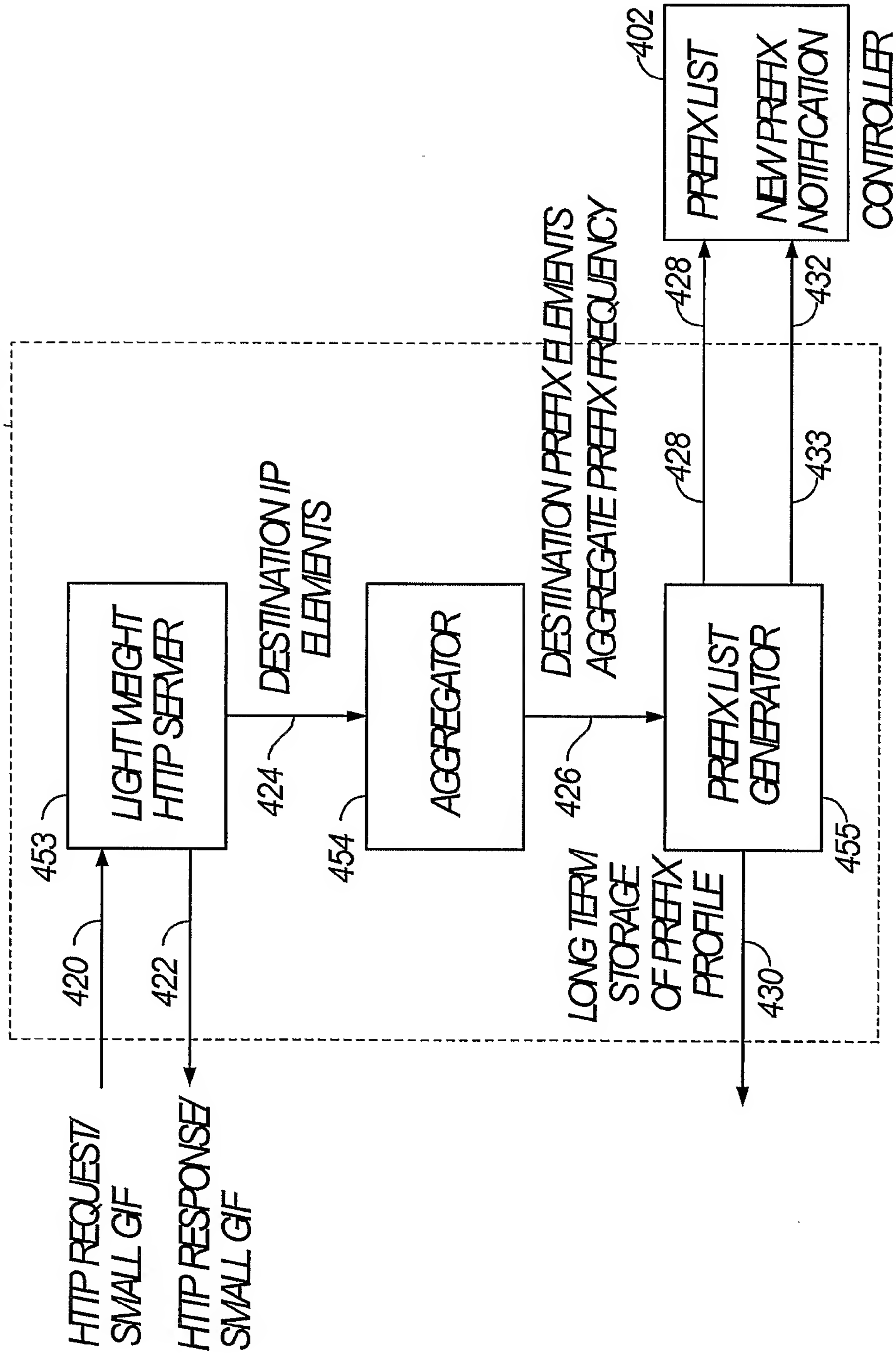

HTTP REQUEST/
SMALL GIF
HTTP RESPONSE/
SMALL GIF
420
422
432
453
LIGHTWEIGHT
HTTP SERVER
424
DESTINATION IP
ELEMENTS
454
AGGREGATOR
426
DESTINATION PREFIX ELEMENTS
AGGREGATE PREFIX FREQUENCY
LONG TERM
STORAGE
OF PREFIX
PROFILE
430
455
PREFIX LIST
GENERATOR
428
433
428
432
402
PREFIX LIST
NEW PREFIX
NOTIFICATION
CONTROLLER

FIG 4

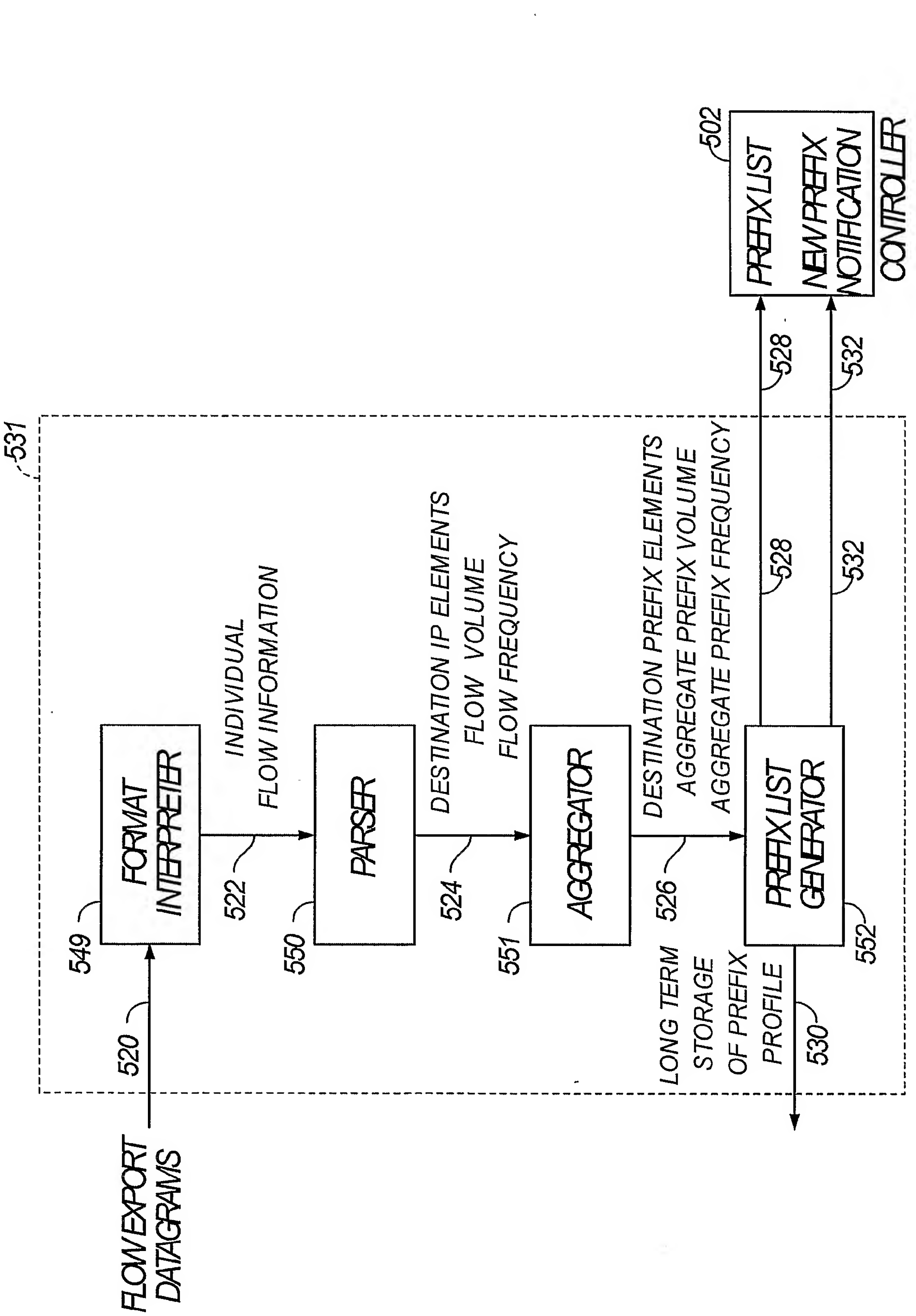

FLOW EXPORT DATAGRAMS
520
531
549
FORMAT INTERPRETER
522
INDIVIDUAL FLOW INFORMATION
550
PARSER
524
DESTINATION IP ELEMENTS
FLOW VOLUME
FLOW FREQUENCY
551
AGGREGATOR
526
DESTINATION PREFIX ELEMENTS
AGGREGATE PREFIX VOLUME
AGGREGATE PREFIX FREQUENCY
552
PREFIX LIST GENERATOR
LONG TERM STORAGE OF PREFIX PROFILE
530
528
532
528
532
502
PREFIX LIST
NEW PREFIX NOTIFICATION
CONTROLLER

FIG. 5

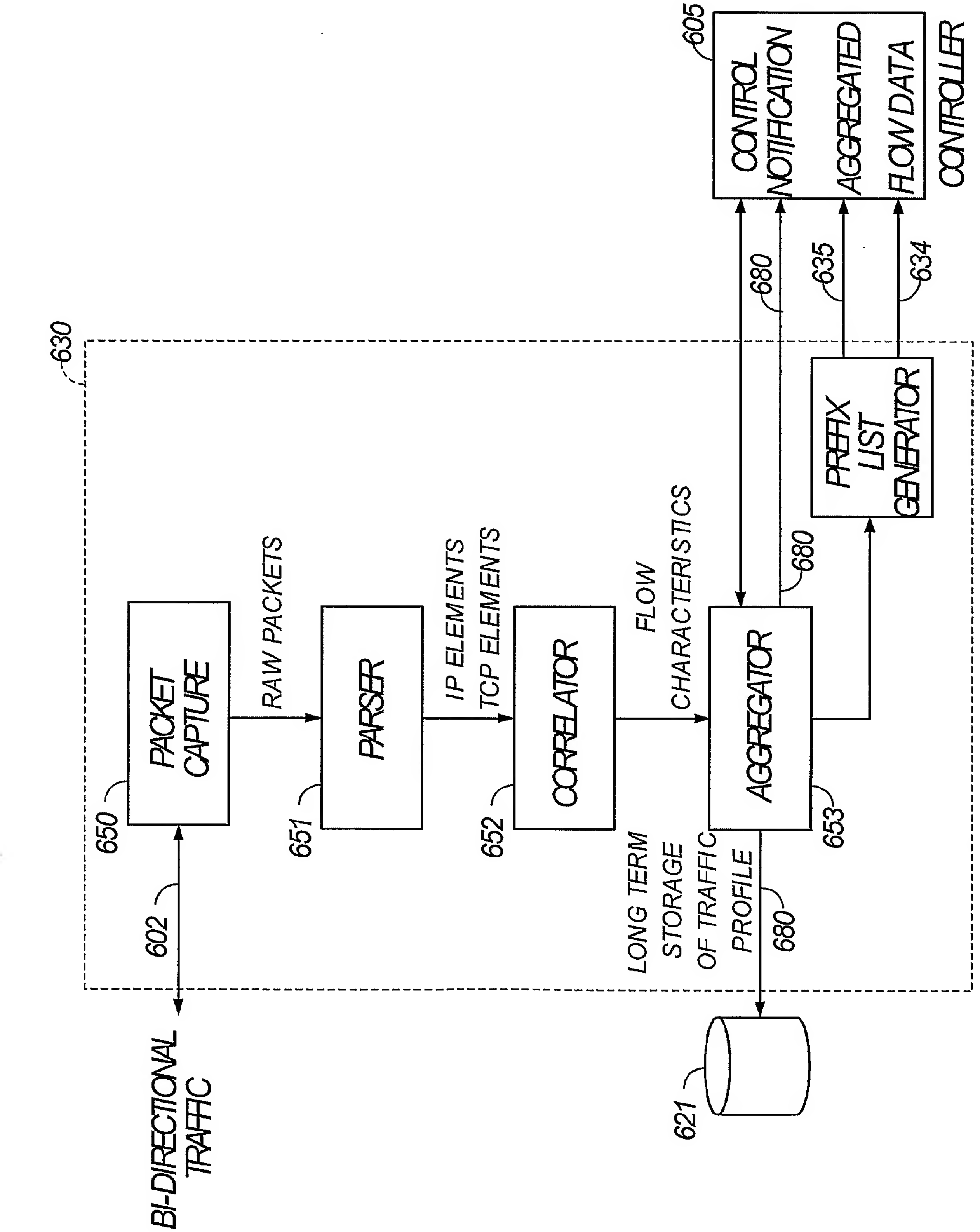

630
BI-DIRECTIONAL TRAFFIC
602
650
PACKET CAPTURE
RAW PACKETS
651
PARSER
IP ELEMENTS
TCP ELEMENTS
652
CORRELATOR
FLOW CHARACTERISTICS
653
AGGREGATOR
680
LONG TERM STORAGE OF TRAFFIC PROFILE
680
621
PREFIX LIST GENERATOR
680
635
634
605
CONTROL
NOTIFICATION
AGGREGATED
FLOW DATA
CONTROLLER

FIG 6

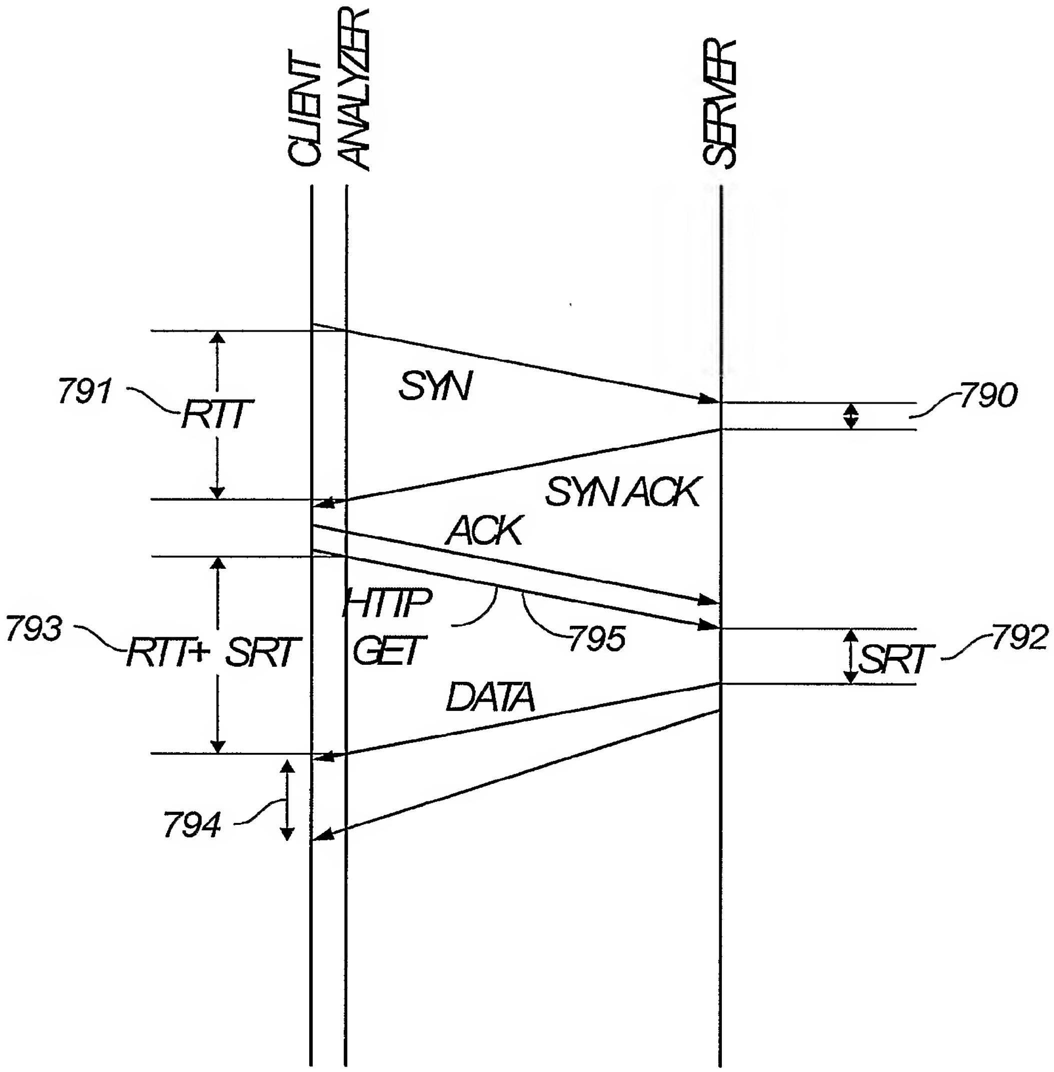

CLIENT
ANALYZER
SERVER
SYN
791
RTT
790
SYN ACK
ACK
793
RTT+ SRT
HTTP
GET
795
SRT
792
DATA
794

FIG. 7

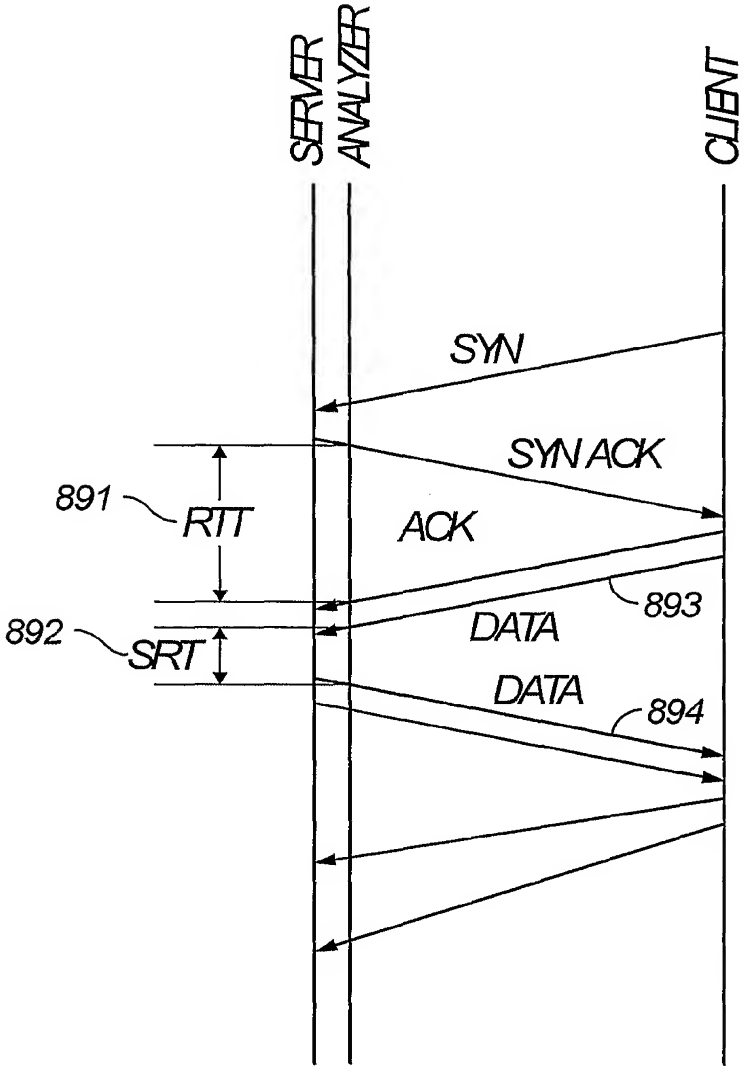

SERVER
ANALYZER
CLIENT
SYN
SYN ACK
891
RTT
ACK
893
DATA
892
SRT
DATA
894

FIG 8

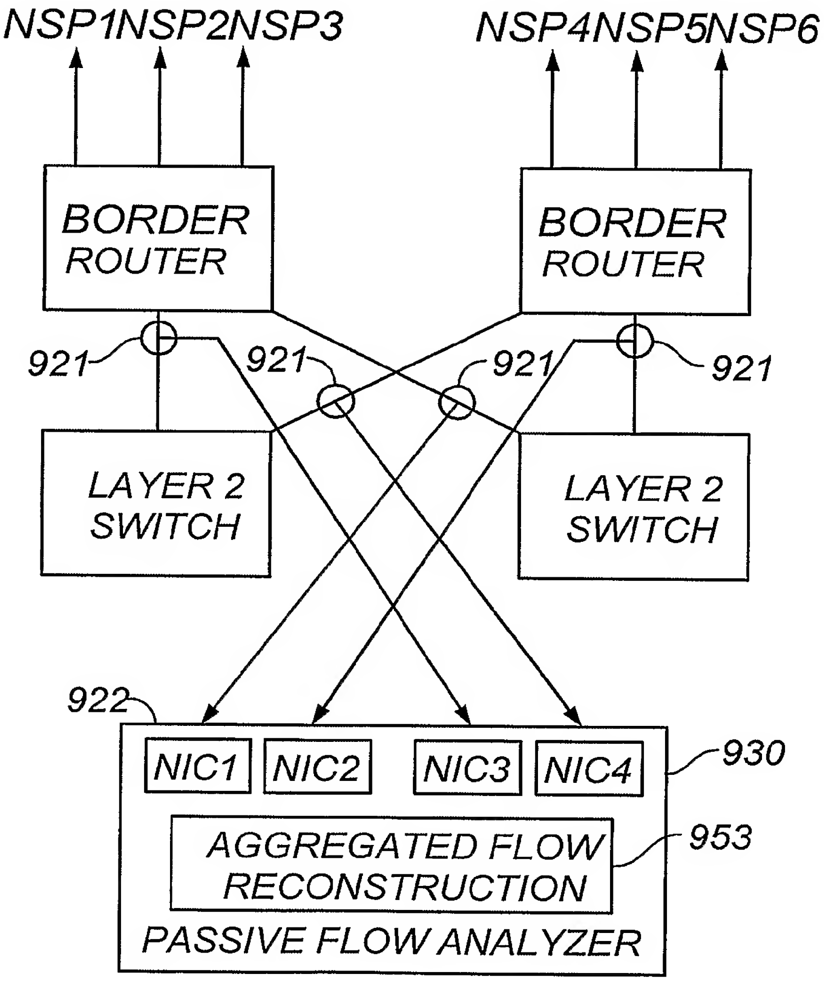

NSP1
NSP2
NSP3
NSP4
NSP5
NSP6
BORDER ROUTER
BORDER ROUTER
921
921
921
921
LAYER 2 SWITCH
LAYER 2 SWITCH
922
NIC1
NIC2
NIC3
NIC4
930
AGGREGATED FLOW RECONSTRUCTION
953
PASSIVE FLOW ANALYZER

FIG. 9

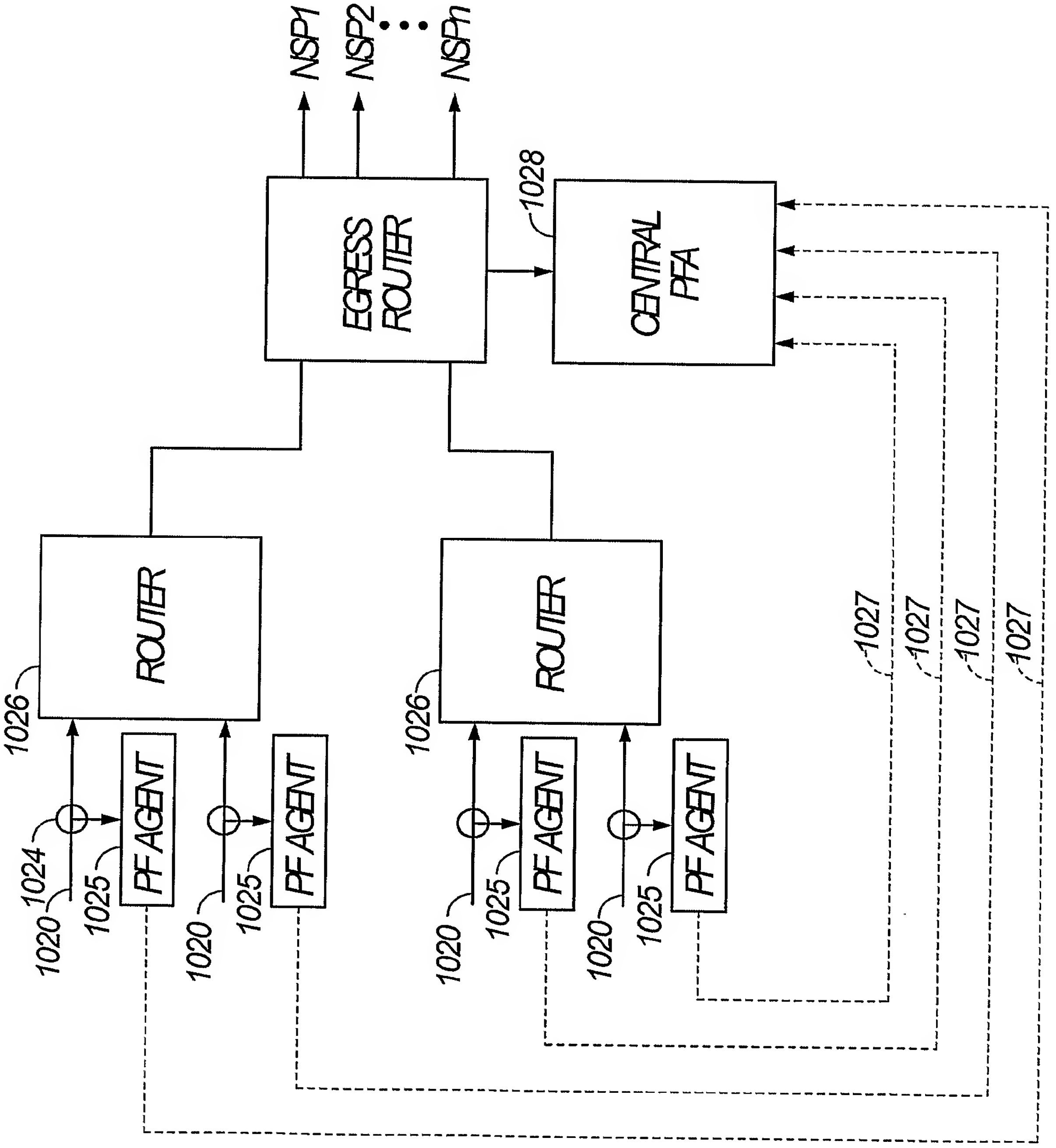

NSP1
NSP2
NSPn
EGRESS ROUTER
1028
CENTRAL PFA
ROUTER
1026
ROUTER
1026
1020
1024
1025
PF AGENT
1020
1025
PF AGENT
1020
1025
PF AGENT
1020
1025
PF AGENT
1027
1027
1027
1027

FIG 10

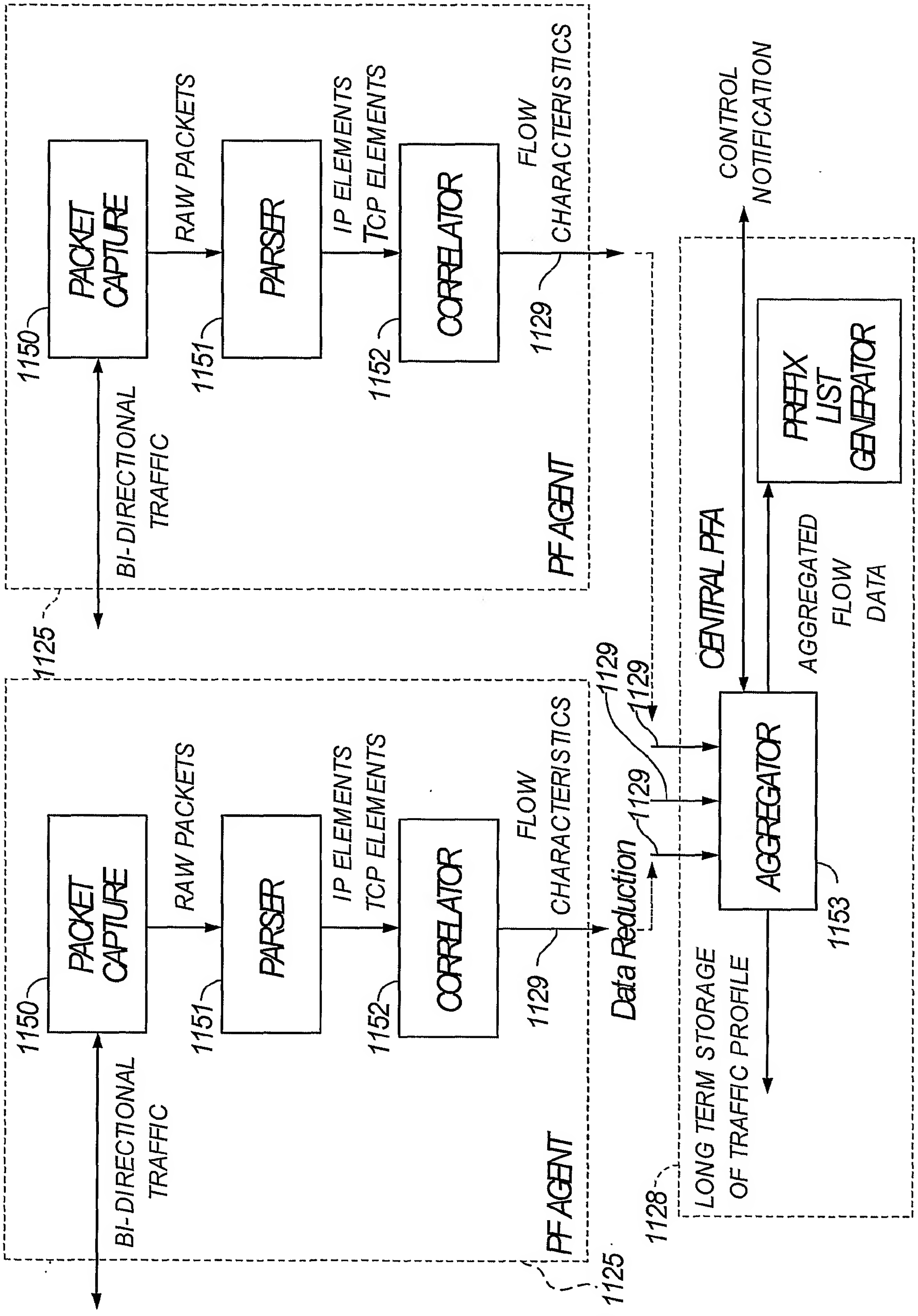

PACKET CAPTURE
RAW PACKETS
PARSER
IP ELEMENTS
TCP ELEMENTS
CORRELATOR
FLOW CHARACTERISTICS
1150
1151
1152
1129
BI-DIRECTIONAL TRAFFIC
PF AGENT
1125
PACKET CAPTURE
RAW PACKETS
PARSER
IP ELEMENTS
TCP ELEMENTS
CORRELATOR
FLOW CHARACTERISTICS
1150
1151
1152
1129
BI-DIRECTIONAL TRAFFIC
PF AGENT
1125
Data Reduction
1129
1129
1129
1129
CONTROL NOTIFICATION
CENTRAL PFA
PREFIX LIST GENERATOR
AGGREGATED FLOW DATA
AGGREGATOR
1153
LONG TERM STORAGE OF TRAFFIC PROFILE
1128

FIG 11

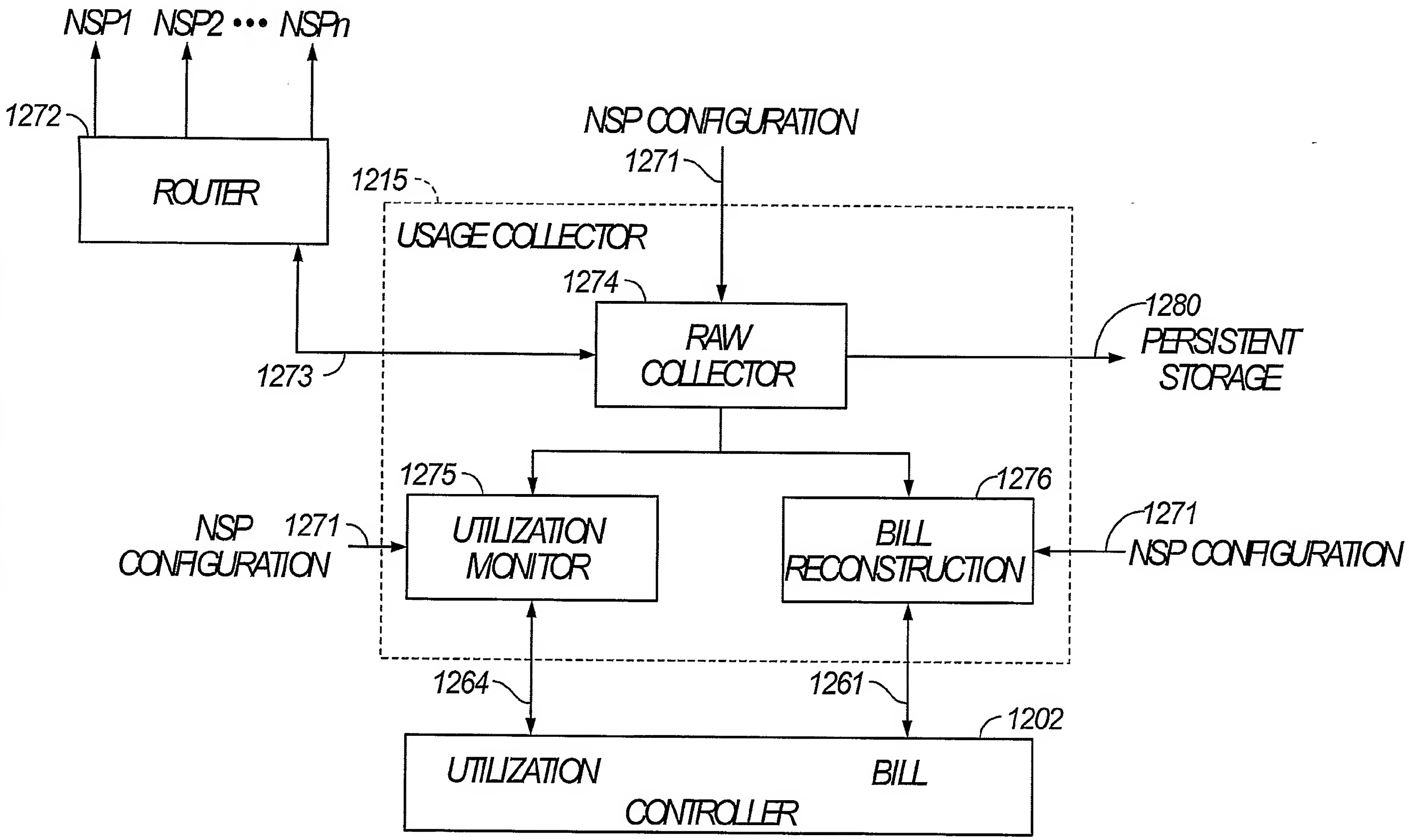

NSP1
NSP2
NSPn
1272
ROUTER
1273
1215
USAGE COLLECTOR
NSP CONFIGURATION
1271
1274
RAW COLLECTOR
1280
PERSISTENT STORAGE
1275
UTILIZATION MONITOR
NSP CONFIGURATION
1271
1276
BILL RECONSTRUCTION
1271
NSP CONFIGURATION
1264
1261
1202
UTILIZATION
BILL
CONTROLLER

FIG 12

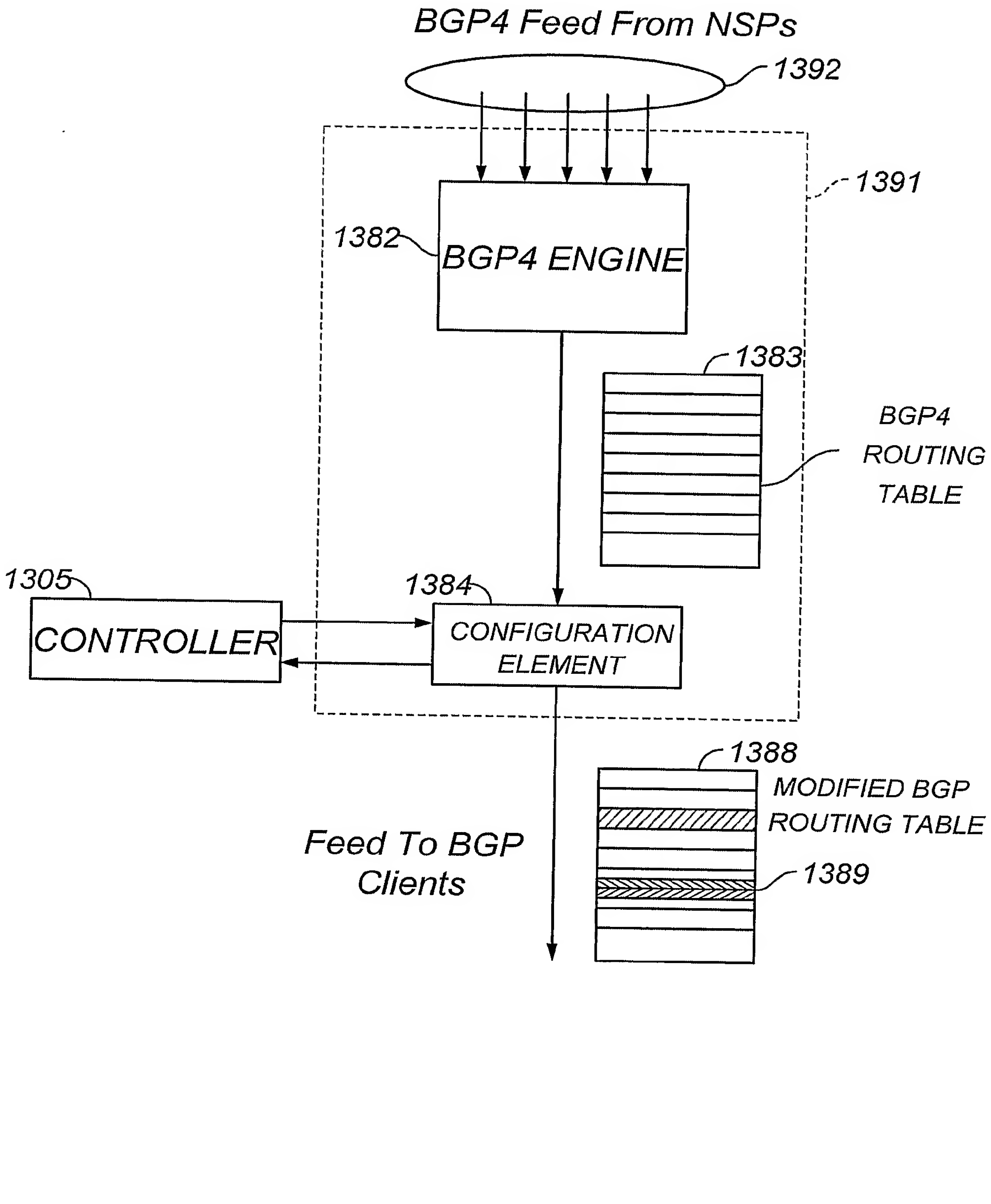

BGP4 Feed From NSPs
1392
1391
1382
BGP4 ENGINE
1383
BGP4
ROUTING
TABLE
1305
CONTROLLER
1384
CONFIGURATION
ELEMENT
1388
MODIFIED BGP
ROUTING TABLE
1389
Feed To BGP
Clients

FIG. 13

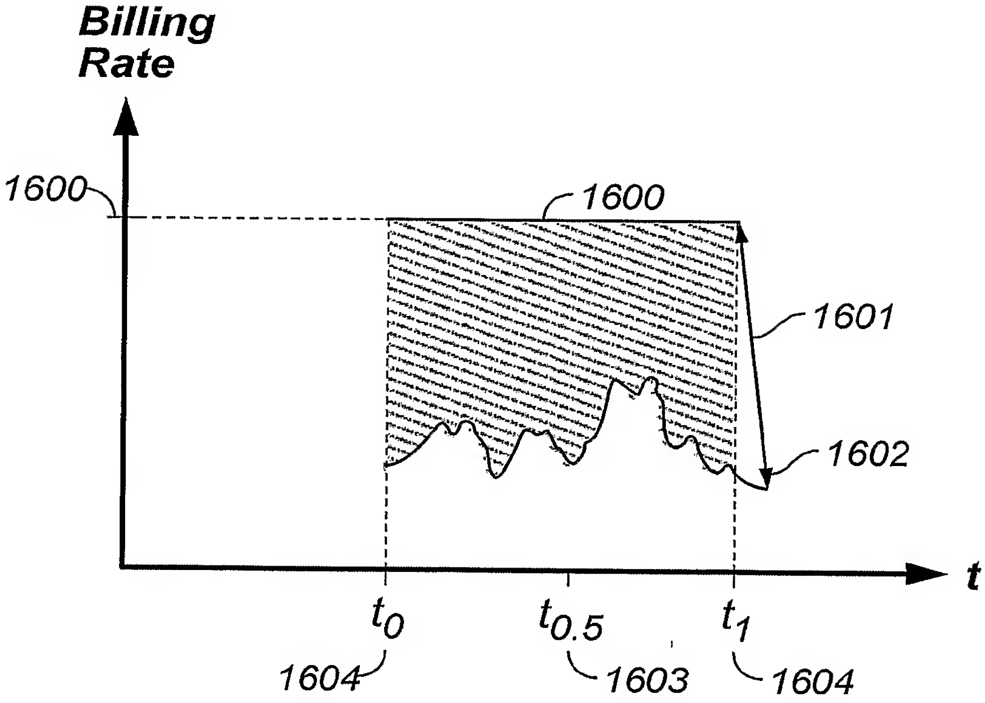


FIG. 14

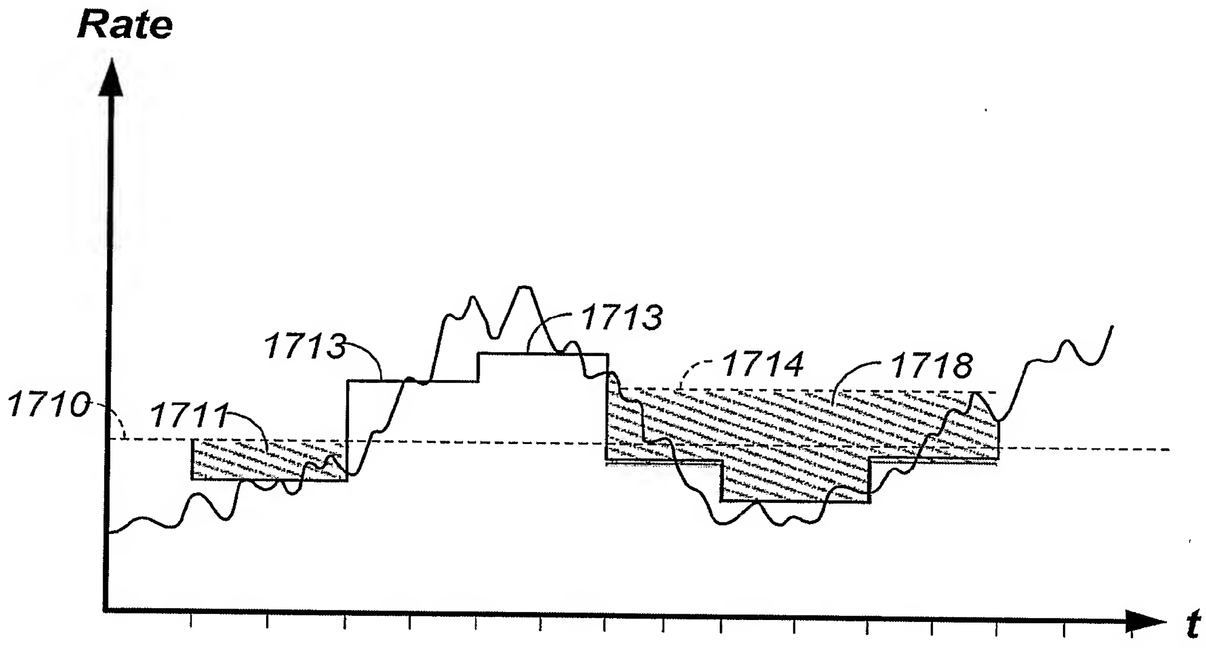


FIG. 15

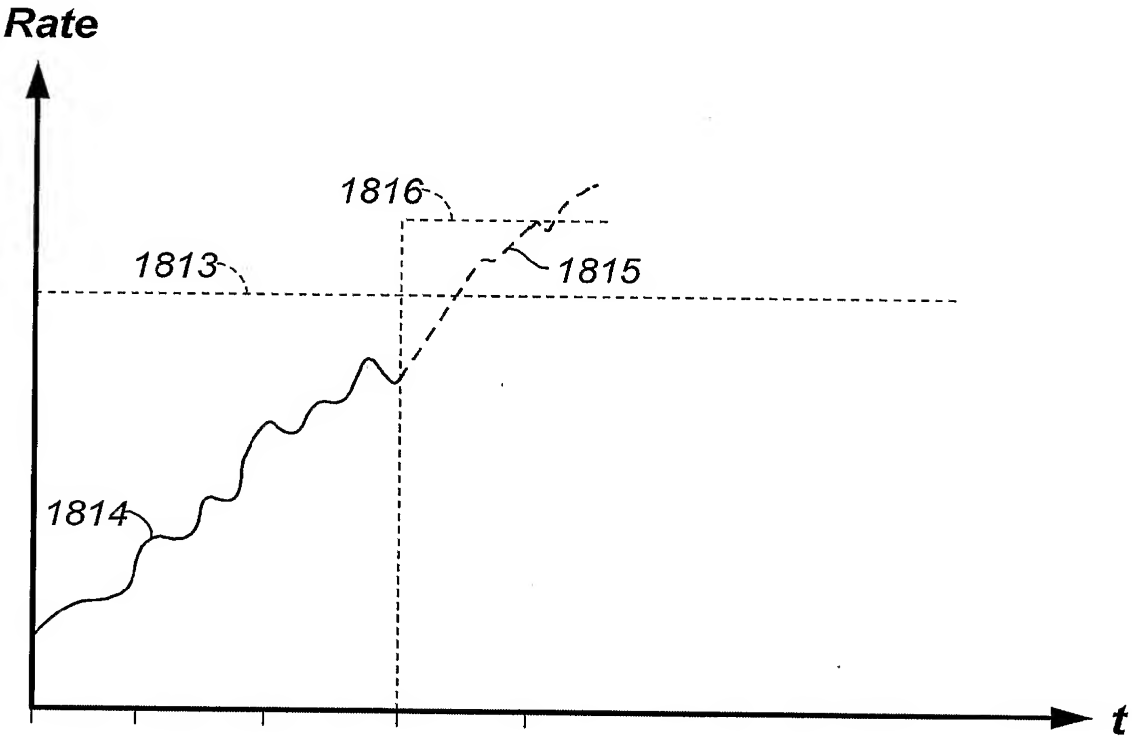
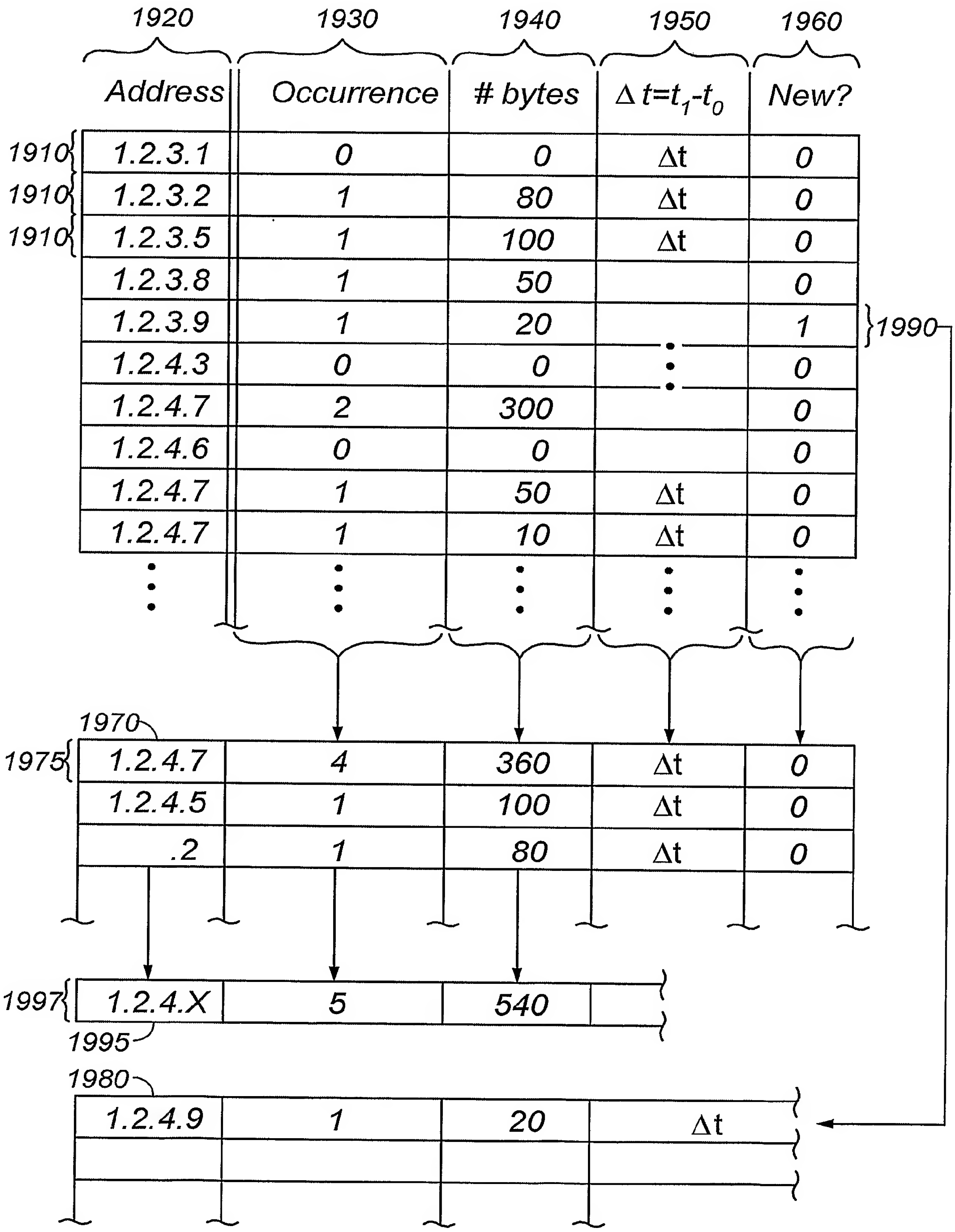


FIG. 16



Address	Occurrence	# bytes	$\Delta t=t_1-t_0$	New?
1.2.3.1	0	0	Δt	0
1.2.3.2	1	80	Δt	0
1.2.3.5	1	100	Δt	0
1.2.3.8	1	50		0
1.2.3.9	1	20		1
1.2.4.3	0	0	⋮	0
1.2.4.7	2	300		0
1.2.4.6	0	0		0
1.2.4.7	1	50	Δt	0
1.2.4.7	1	10	Δt	0
⋮	⋮	⋮	⋮	⋮

1.2.4.7	4	360	Δt	0
1.2.4.5	1	100	Δt	0
.2	1	80	Δt	0

1.2.4.X	5	540	

1.2.4.9	1	20	Δt

FIG. 17

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US02/35158

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : H04L 12/28; H04J 3/14; G01R 31/08; G06F 15/16
US CL : 370/414, 244, 229; 713/201

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 370/414, 244, 229; 713/201

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
NONE

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EAST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y, A	US 6,064,677 A (KAPPLER et al) 16 May 2000 (16.05.2000), column 3, lines 19-67; column 4, lines 1-9.	1-33
Y	US 6,181,679 B1 (ASHTON et al) 30 January 2001 (30.01.2001), column 3, lines 1-67.	1-33
Y	US 6,252,848 B1 (SKIRMONT) 26 June 2001 (26.06.2001), column 3, lines 1-45.	1-33
Y	US 6,226,751 B1 (ARROW et al) 01 May 2001 (01.05.2001), column 8, lines 38-56.	1-33
Y	US 5,870,561 A (JARVIS et al) 09 February 1999 (09.02.1999), column 3, lines 3-47.	1-33

□ Further documents are listed in the continuation of Box C. □ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search 09 December 2002 (09.12.2002)	Date of mailing of the international search report 23 DEC 2002
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703)305-3230	Authorized officer Meng-Ai An Telephone No. (703)305-9669